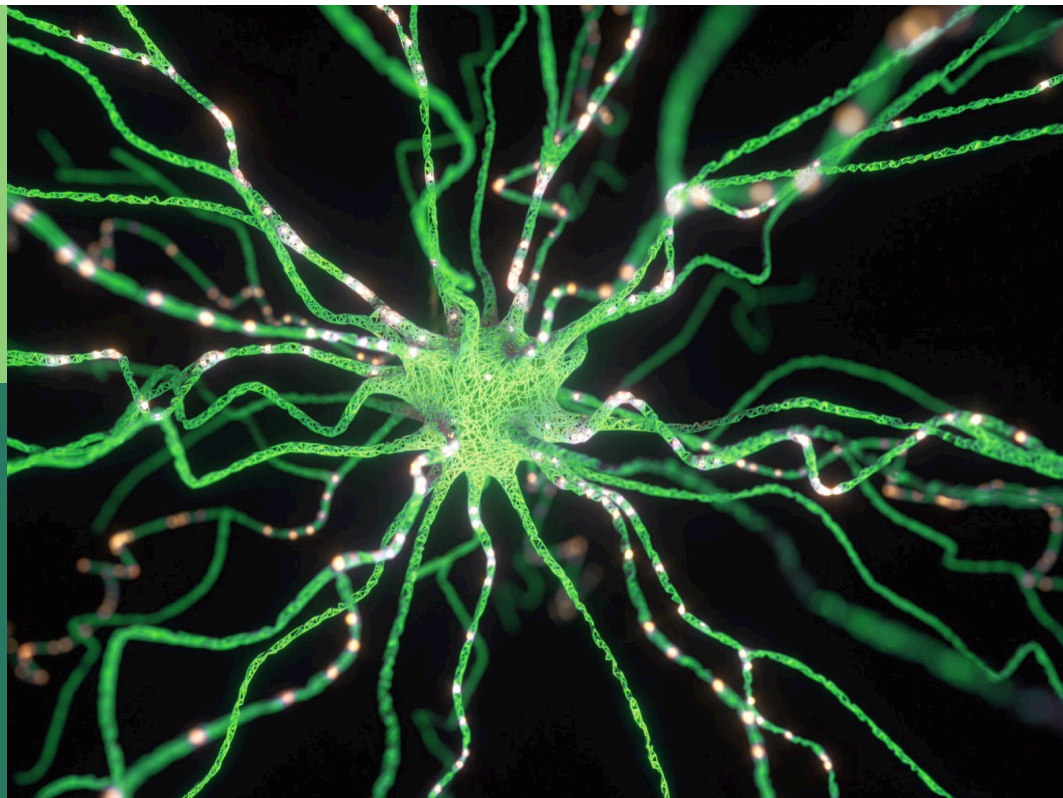
frontiers | Research Topics

# Bio A.I. - from embodied cognition to enactive robotics

**Edited by**
Adam Safron, Inês Hipólito and Andy Clark

**Published in**
Frontiers in Neurorobotics
Frontiers in Psychology
Frontiers in Computer Science



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-3616-2
DOI 10.3389/978-2-8325-3616-2

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Bio A.I. - from embodied cognition to enactive robotics

**Topic editors**

Adam Safron — Johns Hopkins University, United States
Inês Hipólito — Humboldt University of Berlin, Germany
Andy Clark — University of Sussex, United Kingdom

**Citation**

Safron, A., Hipólito, I., Clark, A., eds. (2023). *Bio A.I. - from embodied cognition to enactive robotics*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3616-2

# Table of contents

05 Editorial: Bio A.I. - from embodied cognition to enactive robotics
Adam Safron, Inês Hipólito and Andy Clark

12 The Acquisition of Culturally Patterned Attention Styles Under Active Inference
Axel Constant, Alexander Daniel Dunsmoir Tschantz, Beren Millidge, Felipe Criado-Boado, Luis M Martinez, Johannes Müeller and Andy Clark

25 Enacting Plant-Inspired Robotics
Jonny Lee and Paco Calvo

32 Embodied Intelligence: Smooth Coping in the Learning Intelligent Decision Agent Cognitive Architecture
Christian Kronsted, Sean Kugele, Zachariah A. Neemeh, Kevin J. Ryan Jr and Stan Franklin

52 Situated Neural Representations: Solving the Problems of Content
Gualtiero Piccinini

65 Embodied Object Representation Learning and Recognition
Toon Van de Maele, Tim Verbelen, Ozan Çatal and Bart Dhoedt

84 Self-Concern Across Scales: A Biologically Inspired Direction for Embodied Artificial Intelligence
Matthew Sims

98 Resonance as a Design Strategy for AI and Social Robots
James Derek Lomas, Albert Lin, Suzanne Dikker, Deborah Forster, Maria Luce Lupetti, Gijs Huisman, Julika Habekost, Caiseal Beardow, Pankaj Pandey, Nashra Ahmad, Krishna Miyapuram, Tim Mullen, Patrick Cooper, Willem van der Maden and Emily S. Cross

126 Avoiding Catastrophe: Active Dendrites Enable Multi-Task Learning in Dynamic Environments
Abhiram Iyer, Karan Grewal, Akash Velu, Lucas Oliveira Souza, Jeremy Forest and Subutai Ahmad

149 Social Neuro AI: Social Interaction as the "Dark Matter" of AI
Samuele Bolotta and Guillaume Dumas

159 Goal Oriented Behavior With a Habit-Based Adaptive Sensorimotor Map Network
Felix M. G. Woolford and Matthew D. Egbert

181 Reach Space Analysis of Baseline Differential Extrinsic Plasticity Control
Simon Birrell, Arsen Abdulali and Fumiya Iida

191 Carving Up Participation: Sense-Making and Sociomorphing for Artificial Minds
Robin L. Zebrowski and Eli B. McGraw

202 Grounding Context in Embodied Cognitive Robotics
Diana Valenzo, Alejandra Ciria, Guido Schillaci and Bruno Lara

224 Am I (Deep) Blue? Music-Making AI and Emotional Awareness
Nicholas Novelli and Shannon Proksch

232 The Problem of Meaning: The Free Energy Principle and Artificial Agency
Julian Kiverstein, Michael D. Kirchhoff and Tom Froese

247 Reclaiming saliency: Rhythmic precision-modulated action and perception
Ajith Anil Meera, Filip Novicky, Thomas Parr, Karl Friston, Pablo Lanillos and Noor Sajid

268 Connecting the free energy principle with quantum cognition
Yukio-Pegio Gunji, Shuji Shinohara and Vasileios Basios

282 An enactivist-inspired mathematical model of cognition
Vadim Weinstein, Basak Sakcak and Steven M. LaValle

306 Embodiment enables non-predictive ways of coping with self-caused sensory stimuli
James Garner and Matthew D. Egbert

330 Mind the matter: Active matter, soft robotics, and the making of bio-inspired artificial intelligence
David Harrison, Wiktor Rorot and Urte Laukaityte

349 Using enactive robotics to think outside of the problem-solving box: How sensorimotor contingencies constrain the forms of emergent autononomous habits
Matthew D. Egbert and Xabier E. Barandiaran

372 Small steps for mankind: Modeling the emergence of cumulative culture from joint active inference communication
Natalie Kastel, Casper Hesp, K. Richard Ridderinkhof and Karl J. Friston

frontiers | Frontiers in Neurorobotics

TYPE Editorial
PUBLISHED 14 November 2023
DOI 10.3389/fnbot.2023.1301993

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Jun Tani,
Okinawa Institute of Science and Technology Graduate University, Japan

*CORRESPONDENCE
Adam Safron
asafron@gmail.com

RECEIVED 25 September 2023
ACCEPTED 03 October 2023
PUBLISHED 14 November 2023

CITATION
Safron A, Hipólito I and Clark A (2023) Editorial: Bio A.I. - from embodied cognition to enactive robotics. *Front. Neurorobot.* 17:1301993. doi: 10.3389/fnbot.2023.1301993

COPYRIGHT
© 2023 Safron, Hipólito and Clark. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Editorial: Bio A.I. - from embodied cognition to enactive robotics

Adam Safron[1,2,3,4]*, Inês Hipólito[5] and Andy Clark[5,6]

[1]Center for Psychedelic and Consciousness Research, Department of Psychiatry & Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2]Institute for Advanced Consciousness Studies, Santa Monica, CA, United States, [3]Cognitive Science Program, Indiana University, Bloomington, IN, United States, [4]Kinsey Institute, Indiana University, Bloomington, IN, United States, [5]Department of Philosophy, Macquarie University, Sydney, NSW, Australia, [6]Department of Philosophy and Department of Informatics, University of Sussex, Brighton, United Kingdom

KEYWORDS

**dynamical systems, e-cognition, embodiment, enactivism, active learning, representation, robotics, artificial general intelligence**

Editorial on the Research Topic
Bio A.I. - from embodied cognition to enactive robotics

## Introduction

> "The Brain—is wider than the Sky—
> For—put them side by side—
> The one the other will contain
> With ease—and You—beside—
>
> The Brain is deeper than the sea—
> For—hold them—Blue to Blue—
> The one the other will absorb—
> As Sponges—Buckets—do—
>
> The Brain is just the weight of God—
> For—Heft them—Pound for Pound—
> And they will differ—if they do—
> As Syllable from Sound—"
>
> -Emily Dickinson

If the connections of the human brain were disentangled and placed into a sequence, they would indeed be wider than the sky, being hundreds of kilometers long and likely capable of stretching to the moon and back. If we consider the kinds of intelligence generated by brain-body-environment systems, then such emergent minds may be vaster still in terms of their complex combinatorics, with the pinnacle of expressive power potentially being found in language with its "infinite use of finite means". The field of artificial intelligence and machine learning (AI/ML) seeks to reproduce the powers of biological learners, where we struggle to recapitulate the ways in which even supposedly simple animals demonstrate the ability to respond flexibly to a wide range of situations. In this Research Topic, we were grateful to receive a diverse assortment of articles that address ways in which principles of enactivism and embodied cognition might allow for advances

in AI/ML, potentially without requiring explicit representations, pre-specified algorithms, or centralized control structures. In what follows, we briefly summarize these contributions, highlight some potential implications, and end with a discussion of potential ways forward for AI/ML and cognitive science more generally.

## Summary of contributions and commentary

*Please note that while we use the author's own words where possible, we strongly encourage interested readers to reference the original articles.*

In "*The acquisition of culturally patterned attention styles under active inference*", Constant et al. present simulations of visual foraging based on active inference, demonstrating the acquisition of attention styles patterned according to cultural artifacts that drive perception, action, and learning. This paper compellingly shows how material culture may both drive and be driven by human thought and by the building and rebuilding of patterns of attention.

In "Enacting plant-inspired robotics", Lee and Calvo suggest plants as a holistic source of inspiration for soft robotics in terms of their non-centralized, modular architectures and highly plastic phenotypes. In contrast with notions of autonomy based on the independent operability of systems over an observation window, plants and other living organisms exhibit a stronger form of autonomous functioning in terms of needing to support self-production dynamics that create distinctions between themselves and the "domain of interactions that maintain the conditions of viability for the system". They further suggest that the field of "growbots" could be advanced if those systems took a more active role in acquiring sources of matter and energy for the sake of self-preservation.

In "*Carving up participation: sense-making and sociomorphing for artificial minds*", Zebrowski and McGraw argue that properly understanding social cognition requires a greater appreciation of the nature of interactions involving participatory sense-making (PSM). "Sociomorphing" is proposed as a means of distinguishing between living sense-makers and artificial systems, potentially allowing for the gradual incorporation of AIs into contexts involving initially asymmetric degrees of sociality. PSM and sociomorphing are suggested to provide not only a basis for social robotics but also a potentially robust framework for developing increasingly advanced AIs with general intelligence.

In "*Embodied object representation learning and recognition*", Van de Maele et al. show how robotics can be informed by considering the ways in which biological agents achieve scene understanding for adaptive object manipulation and navigation capabilities by leveraging active interactions with the world from their first encounters with novel situations. Taking inspiration from theories of neuroscience in which neocortical columns build predictive models about objects within allocentric reference frames, the authors introduce a Cortical Column Network (CCN) architecture. In CCNs, each object category is represented in its own reference frame by learning a generative model over expected/predicted transformations in pixel space, given actions. CCN ensembles vote on their respective beliefs regarding candidate object categories, which results in the creation of novel CCNs when classification likelihoods are too low. This architecture is further validated in simulation environments, with classification improving as agents gather more evidence (with self-supervised active learning) and choose actions in ways that afford reaching preferred observations/destinations.

In "Grounding context in embodied cognitive robotics", Valenzo et al. describe how autonomous machines may be augmented with greater behavioral flexibility by providing systems with a "global context" that integrates agent-related, environmental, and task-related information. Through the interaction of these core elements, agents are capable of (1) selecting self-relevant tasks on the basis of current and anticipated future needs (for learning and mastering contingencies), (2) performing tasks with continuous performance monitoring, and (3) abandoning unsuccessful tasks based on overall prediction errors during situated action cycles. With respect to prediction-error monitoring, the rate of reduction is taken as an index of overall performance success, evoking emotions that both function as driving elements for autonomous behavior and are also shaped by the interactions of core elements of global context processing.

In "*The problem of meaning: the free energy principle and artificial agency*", Kiverstein et al. describe how biological agents solve the "problem of meaning", by acting in ways that express sensitivity to context-dependent relevance. Drawing on common principles of mind-life continuity and enactivist cognitive science, the authors argue that robustly autonomous agents require stable, self-sustaining patterns of sensorimotor interaction to ground values, norms, and goals as they encounter different (and differently) meaningful environments. The authors further discuss relationships between enactivism and the FEP, including the challenge that these perspectives are fundamentally incompatible, with biological systems exhibiting historical path-dependent learning but with free-energy-minimizing agents severing this historicity. Such FEP agents also show a lack of the "interactional asymmetry" present in enactivist accounts of autonomy. In addition to addressing these challenges, it is suggested that rather than fundamental incompatibility, the FEP needs enactivism for the problem of meaning, and enactivism needs the FEP for precise formal modeling of the necessary constituent factors for realizing agency.

In "*Avoiding catastrophe: active dendrites enable multi-task learning in dynamic environments*", Iyer et al. introduce a neural network architecture for enhancing the embodied systems to operate in dynamic environments while flexibly adapting to changing task contexts and continuously learning without catastrophic forgetting/interference. This is achieved by incorporating active dendrites and sparsity-promoting local inhibitory systems, so dynamically constraining and routing information in a context-specific manner. The architecture is tested on several benchmarks, including a multi-task reinforcement learning environment in which agents must solve a variety of manipulation tasks (cf. meta-learning), in addition to a continual learning setup in which task predictions change over the course of training (cf. reversal learning). In both simulations, the architecture developed overlapping yet distinct sparse subnetworks that mediated the fluid adaptation to multiple tasks with

minimal forgetting, providing (for the first time) a demonstration of high performance with respect to both multitasking and continual learning.

In "*Social neuroAI: social interaction as the "dark matter" of AI*", Bolotta and Dumas introduce a three-axis framework for social learning in biologically-inspired AI, informed by FEP-AI: (1) brain-inspired models of cognitive architectures, such as global workspace and attention schema theories, that bridge individual and social intelligence; (2) dynamical systems perspectives for handling the inherently time-dependent nature of cognition; (3) embodiment as a source of sophisticated communicative signals. These social interactions are essential elements of advanced cognitive ability yet remain under-explored in AI, constituting the "dark matter" with respect to attempts to understand human(imal)-like intelligence. In light of this gap in our understanding, the authors review the role of social learning in cognitive development and the emerging field of "Social NeuroAI."

In "*Goal-oriented behavior with a habit-based adaptive sensorimotor map network*", Woolford and Egbert present a habit-based robot controller model that draws on enactivist principles to realize agency via an adaptive sensorimotor map (ASM) network architecture. ASM networks provide platforms for experimental investigation that combine (1) mechanisms for generating continuous motor activity as a function of historical trajectories and (2) evaluative mechanisms that reinforce or weaken those trajectories as a function of their support for the structure of higher-order sensorimotor coordination. The authors deploy these adaptive networks in a minimal cognition task involving object discrimination, demonstrating how an individual robot could learn through a combination of exploratory/random movements and repetition of successful historical trajectories of sensorimotor coordination (cf. motor babbling). These robots display learning without explicit representational mechanisms or extraneous fitness variables but rather adapt according to the internal requirements of the action-generating mechanisms themselves.

In "*Embodied intelligence: smooth coping in the learning intelligent decision agent cognitive architecture*", Kronsted et al. describe how skillful actions may become habituated and ingrained through experience, thereby placing less stress on cognitive load relative to considered and deliberative thought and action (e.g., walking, driving, skiing, playing music, short-order cooking). Smooth coping behaviors appear to be automatized in that they are rapid and lacking in reflection, corresponding to Hurbert Dreyfus' description of Heideggerian phenomenology involving "mindless" absorption in action and being in a state of flow. However, pragmatists such as John Dewey et al. suggest that intelligent flexibility is built into smooth coping in ways that make it distinct from automatization. The authors detail a conceptual model of smooth coping using the Learning Intelligent Decision Agent (LIDA) system, informed by the Global Workspace Theory of Consciousness, and argue that sequences of automatized actions are intermittently interspersed with skillful and flexible adjustment by consciously-mediated action selection (via dorsal stream processes). An Automatized Action Selection sub-module is introduced into LIDA to demonstrate these principles within a hybrid architecture that allows for a synergistic combination of both enactivist couplings and explicit representation for the sake of more skillful conscious control of behavior.

In "*Situated neural representations: solving the problems of content*", Piccinini argues that situated approaches to mind based on embodiment, embedding, enaction, and affect (with extension not being relevant to their discussion) are deeply intertwined with neural representation, with such a computational approach "[requiring] embodiment, embedding, enaction, and affect at its very core." Additionally, situatedness is suggested to be necessary to describe the adaptive shaping of computations in ways that (1) construct representations with original semantic content, (2) automatically coordinate neural vehicles with representational content, (3) allow content to be causally efficacious, (4) allow content to be sufficiently determinate to be meaningful/useful to systems, (5) allow representation of distal stimuli, and (6) allow for the possibility of misrepresentation.

In "*An enactivist-inspired mathematical model of cognition*", Weinstein et al. outline an enactivist-compliant mathematical framework for natural and artificial cognitive systems that do not attribute contentful symbolic representations to agents but instead model nervous systems, bodies, and environments as "an inseparable part of a greater totality". Sensorimotor systems are considered to be special cases of (potentially labeled) "transition systems" with connections to deterministic automata. Minimal sufficient requirements are also suggested for the property of "sufficiency", including optimal attunement of an organism to its environment with sufficient history information spaces.

In "*Using enactive robotics to think outside of the problem-solving box: how sensorimotor contingencies constrain the forms of emergent autonomous habits*", Egbert and Barandiaran suggest that AI ought to take inspiration from the "precarious, self-maintaining organization of living systems". They demonstrate how robots controlled by an iterative Deformable Sensorimotor Medium can realize the spontaneous emergence of an organized ecology of habits capable of re-enacting adaptive behaviors, with habits formed within modalities having relatively greater similarity to habits across modalities (similar to observations for biological systems). These findings are further discussed in terms of their relevance to sensorimotor contingency theory, adaptationist and structuralist explanations in biology, and the potential limitations of functionalist problem-solving approaches to AI.

In "*Reach space analysis of baseline differential extrinsic plasticity [(DEP)] control*", Birrell et al. introduce a learning rule studied in the context of goal-free simulated agents that produce environmentally aware behaviors. They further extend this mechanism to intentional behavior to determine whether "short-circuited DEP" can generate desired trajectories in a robotic arm via simple open-loop control, with transient and limit cycle dynamics explored in experiments involving target reaching and circular motions.

In "*Resonance as a design strategy for AI and social robots*", Lomas et al. explore the relationships between the physical mechanisms of resonance and human experience, with consideration for enhancing those (potentially highly impactful) experiences within human-robot interactions. The authors discuss resonance as a cultural and scientific metaphor and review "sympathetic resonance" as a physical mechanism (including

synchronization and rhythmic entrainment) and "design strategy" for shaping interactions between human and non-human systems.

With "*Self-concern across scales: a biologically inspired direction for embodied artificial intelligence*", Sims focuses on a foundation for intelligence for all biological systems that reflects the existential task of continued viability. Self-concern is introduced as "a property of a complex system that describes its tendency to bring about states that are compatible with its continued self-maintenance", and a potential means of recapitulating the power (and principles) of human-like intelligence in artificial systems.

With "*Mind the matter: active matter, soft robotics, and the making of bio-inspired artificial intelligence*", Harrison et al. argue for limitations in the realizability of cognitive phenomena such as memory, learning, goal-directedness, and decision-making. That is, the authors describe how cognition is deeply intertwined with its materiality and corporeality and suggest that progress in AI may require treating the underlying material, living processes as more than mere "hardware" that can be abstracted over without consideration for the soft, active, and plastic details of the particular mechanistic realizers. In short, "the matter matters for cognitive form and function." With "multiple realisability 2.0", materiality enables, mediates, and constrains cognition, with precarious conditions for existence being essential for understanding how autonomous systems value, engage, and interact with their environments with a goal-directedness grounded in existential needs of survival, persistence, and reproduction.

In "*Reclaiming saliency: rhythmic precision-modulated action and perception*", Anil Meera et al. characterize the nature of visual attention and saliency and how standard accounts based on mutual information between current visual information and estimated causes fail to consider the circular causality linking perception and action (including decisions as to where to sample next, given current beliefs). From this perspective, salience is defined as an active inferential process that relies on the basic principles of uncertainty minimization and rhythmic scheduling and attention: precision control, or the confidence with which beliefs can be updated, given sampled sense data. Alternatively phrased, salience is related to uncertainty minimization, underwriting the selection of future sense data, and attention is related to rhythmic precision modulation. Numerical experiments are provided to demonstrate advantages for state and noise estimation, as well as system identification and action selection for informative path planning.

In "*Embodiment enables non-predictive ways of coping with self-caused sensory stimuli*", Garner and Egbert demonstrate how sensory attenuation for self- (relative to externally-) caused stimuli can be explained enactively. This is contrasted with classical explanations of these phenomena based on efference copies, wherein motor commands are accompanied by copies of signals that predict the likely sensory consequences of that activity, which are then subtracted from the actual sensory input. Genetic algorithms are used in this work to investigate when non-predictive solutions might be viable, which in the simple systems tested involved modifying paper to shape or avoid self-caused sensory inputs (rather than predicting and filtering them out) and sometimes leveraging these self-caused inputs for greater control, all without the need for an explicit internal model.

In "*Am I (Deep) Blue? Music-making AI and emotional awareness*", Novelli and Proksch provide a review of the applications of AI to creative and emotional artistic endeavors, focusing on musical composition. The authors suggest limitations of systems rooted in current AIs that lack "thoroughly embodied, interoceptive processes associated with the emotional component of music perception and production". The authors' review presents attempts to combine the impressive power of modern generative models with more human-like emotional/interoceptive processing.

In "*Connecting the free energy principle with quantum cognition*", Gunji et al. outline a potential conflict between FEP-AI and quantum cognition. While free energy minimization leads to a Boolean lattice of classical logical propositions, quantum cognition leads to an orthomodular lattice of quantum logical propositions. Excess Bayesian inference is introduced, with binary relations transformed from a distribution of the joint probabilities via rough-set lattice techniques.

In "*Small steps for mankind: modeling the emergence of cumulative culture from joint active inference communication*", Kastel et al. provide a compelling and testable deep active inference formulation of social behavior and simulations of cumulative culture. Cultural transmission is cast as a bi-directional communication process that induces particular convergences (via generalized synchrony) between the belief states of interlocutors. Social/cultural exchange is further cast as a process of active inference, equipping agents with choices regarding who to engage with as communication partners, thus inducing trade-offs between confirmation of current beliefs and exploration of social environments. Cumulative culture emerges from the dynamics of belief updating, with equilibria manifesting as segregation into groups whose belief systems are actively sustained through selective, uncertainty-minimizing, dyadic exchanges. Finally, the nature(s) of these emergent equilibria crucially depend on the precision-weighting of each individual's generative model of their encultured niches.

## Conclusion

Across these contributions, we can see a broad range of views on what it means for a system to be biologically inspired, many of which are still neglected in machine learning. For example, people are increasingly interested in enhancing large language models with "multimodality" and potential grounding via simulation environments (Driess et al., 2023; Yin et al., 2023). However, approaches that attempt to take on enactivist insights are rare, with business-as-usual oftentimes assuming that we might be able to rely on achieving new emergent capabilities with sufficient scaling (Silver et al., 2021). This is in contrast to what might be suggested from fields such as developmental social robotics, which emphasize the conditions for bootstrapping (and grounding) robust and flexibly generative models of systems that "grasp" an organism's meaningful interactions with the environment (Dreyfus, 2007; Tani, 2016; Kolchinsky and Wolpert, 2018; Linson et al., 2018; Bisk et al., 2020; Safron, 20211; Hipólito et al., 2023).

From a radically embodied perspective, one might argue that the entire field of cognitivist deep learning is on shaky foundations by virtue of needlessly appealing to the literal sense of the mind-machine metaphor, i.e. to minds as literal information processors

(van Gelder, 1990; Van Gelder, 1995; Hutto and Hipólito, 2021; Beckmann et al., 2023). In their view, because computation and information processes cannot be found "in the wild" independent of human (scientific) practices, the literal sense of the analogy pushes toward a rudimentary view of natural intelligence (even if operationally useful in some circumstances). However, we believe that a more ecumenical approach may be called for if we relax some of the usual assumptions that accompany these more cognitivist notions, which may perhaps be made more powerful (and flexible) when re-represented in more enactivist terms. For example, one may think of a diverse range of scientific representations for understanding biological intelligence without necessarily endorsing that the target being represented entails the ontological properties of the model (Candadai and Izquierdo, 2020; Constant et al., 2020). These include (but are not limited to) the following models (of representation/modeling-like phenomena):

1. Implicit "representation" and generalized stigmergic auto-encoding of action-perception cycles via distributed attractor dynamics over likely patterns of enaction with information continuously with/offloaded into the environment in an extended mind sense (Clark and Chalmers, 1998; Pfeifer and Bongard, 2006; Heylighen, 2016).
2. Partially disentangled features in shared latent workspaces (Bengio, 2017; Thomas et al., 2017, 2018)—possibly centered in posteromedial and lateral parietal cortices (Safron, 2021a)—potentially describable as reduced-dimension manifolds over which neuronal activity evolves (Ji et al., 2023).
3. Predictive modeling of the likely homeostatic consequences of different system-world states by subcortical structures that ground all cognition in the preconditions for successful life management and reproduction (Damasio, 2012; Safron, 2021b; Solms, 2021), thus coupling the individual to phylogenetic (meta-)learning (Campbell, 2016; Ramstead et al., 2018; Botvinick et al., 2019; Safron, 2019; Wang, 2021).
4. Predictive modeling (and thereby control) of these system-world estimates by value-canalized striatal-cortical loops could be understood as conditioning these percepts/concepts on likely patterns of enaction. At hierarchically lower levels, these could take the form of softly assembled coalitions of forward models (cf., amortization and planning as inference) (Botvinick and Toussaint, 2012; Kaplan and Friston, 2018). At intermediate levels of abstraction, these could take the form of (experienceable) patterns of embodied simulation and the structuring of perception by relevant affordances (Cisek, 2007). At higher levels, these could take the form of (not directly experienceable) patterns of recurrent activity (or reservoirs), whose bifurcations/tensors could flexibly parameterize likely patterns of enaction with capacities for evaluating multiple policies (Tani, 2016).
5. Re-representation of these features in the spatiotemporal trajectories of the hippocampal/entorhinal system (Blouw et al., 2016; Whittington et al., 2020; George et al., 2021; Safron et al., 2021; Bengio et al., 2022; Dumont et al., 2023), so allowing for orchestration of large-scale dynamics by likely state transitions for the overall agentic system through time-space, potentially affording some of the kinds of graphical representations associated with "good-old-fashioned AI" and symbolic cognitive science (Gentner, 2010; Crouse et al., 2020).
6. Local object models (Kosiorek et al., 2019; Van de Maele et al.), which would be consistent with characterizations of cortical columns as types of transformers, or Numenta's "1000 brains theory" (Hawkins, 2021). While it is questionable whether every cortical column entails full allocentric object modeling capabilities (Safron et al., 2021), this may be the case for local "modules" that are capable of achieving sufficient degrees of functional closure with respect to being able to inform and be informed by action-perception cycles on the timescales of their formation (e.g., whisker barrels, but not ocular dominance columns). This is an example of how seemingly cognitivist models of mental phenomena involving "representation" may heavily depend on an understanding of enactivist principles to accurately characterize the specific details of the operation.
7. Re-representation of these features through symbolic/linguistic capacities (which are themselves realized as probable patterns of enaction for partially expressed motor sequences/grammars), thus allowing for cognition to be structured/stabilized/expanded according to the combinatorics of syntactic language with its "infinite use of finite means". By affording multi-level recursive self-referential self-modeling, an additional set of strange-loop-involving (Hofstadter, 2007) virtual machines is placed on top of "cognitive" hierarchies, thereby expanding "cognitive light cones" to indeed be "wider than the sky."—For a preliminary discussion, see Friston et al. (2023).

In this non-exhaustive list of methodologies, it may be possible to find an inclusive, potentially synergistic, and scientifically valuable middle ground between seemingly incompatible theories on the understanding of the mind. This effort is illustrated in the diverse articles in this collection, ranging from discussions of the centrality and power of morphological computation to demonstrations of the promise of biologically-inspired neural architectures.

It is worth noting that this more ecumenical stance still requires criticality, as we would also caution against assuming that adding seemingly biological features to a system will necessarily improve its intelligent/adaptive functioning. This cautioning may be especially timely in light of trends in AI/ML that attempt to project future gains in performance based on a combination of apparent "laws" of ability scaling with computation, especially when combined with analogies regarding human brains as "neural networks". Of course, brains are indeed types of neural networks, but they also have multiple heterogeneous subsystems, which, taken together, create a control architecture for embodied agents embedded in environments in which they pursue valued goals, usually developed (or trained) in the context of intelligently-structured socioemotional learning curricula (Tomasello, 2014; Veissière et al., 2019; Safron, 20211). As such, attempts to reduce the sophistication of cognition to a "master algorithm" are likely doomed to failure.

Moreover, a substantial amount of intelligent functioning may be realizable via the morphological "computation" enabled by intelligently designed body plans and their physical reactive dispositions. Indeed, this kind of "offloading" of computational challenges onto (or into) bodies and environments is precisely what we would expect from predictive processing systems as they attempt to achieve adaptive functioning with maximal efficiency. While "explaining away" prediction errors via dynamics closer to primary modalities requires fewer neuronal transactions than leveraging more complicated models, the energetic savings (of minimizing cybernetic entropy) are even greater still if prediction errors never enter nervous systems in the first place because they have been eliminated via (en)active inference (Ramstead et al., 2019). It follows, we believe, that the most fruitful meta-prior/over-hypothesis for enactivism-informed cognitive science would be that when it comes to trying to understand the sources of biological intelligence, one should begin with observational behavior and how cognition emerges from a system's interaction with its context-sensitive environment.

We are grateful to have had the opportunity to help bring together this collection on the diverse ways in which embodiment and environmental interactions provide foundations for cognition, across multiple scales. While it may still be debated the precise ways in which systems must be embodied in order to realize which degrees (and kinds) of intelligence, we would even go so far as to conclude with the maxim: "no body, never mind." Or, in the words of the great late poet Mary Oliver: "The spirit likes to dress up like this: ten fingers, ten toes, shoulders, and all the rest· · · It could float, of course, but would rather plumb rough matter. Airy and shapeless thing, it needs the metaphor of the body· · · it needs the body's world· · · to be understood, to be more than pure light that burns where no one is-so it enters us· · · lights up the deep and wondrous drownings of the body like a star" (Oliver, 1986).

## Author contributions

AS: Writing—original draft, Writing—review and editing. IH: Writing—original draft, Writing—review and editing. AC: Writing—original draft, Writing—review and editing.

## Funding

AC was supported by European Research Council, Synergy Grant (XSCAPE) ERC-2020-SyG 951631.

## Acknowledgments

We would like to thank Victoria Klimaj for her invaluable support in assisting with this Research Topic.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Beckmann, P., Köstner, G., and Hipólito, I. (2023). An alternative to cognitivism: computational phenomenology for deep learning. *Minds Mach.* 33, 397–427. doi: 10.1007/s11023-023-09638-w

Bengio, Y. (2017). *The Consciousness Prior*. arXiv:1709.08568 [Cs, Stat]. Available online at: http://arxiv.org/abs/1709.08568 (accessed September 24, 2023).

Bengio, Y., Deleu, T., Hu, E. J., Lahlou, S., Tiwari, M., and Bengio, E. (2022). *GFlowNet Foundations (arXiv:2111.09266)*. arXiv.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., et al. (2020). *Experience Grounds Language.* arXiv:2004.10151 [Cs]. Availabe online at: http://arxiv.org/abs/2004.10151 (accessed September 24, 2023).

Blouw, P., Solodkin, E., Thagard, P., and Eliasmith, C. (2016). Concepts as semantic pointers: a framework and computational model. *Cogn. Sci.*, 40, 1128–1162. doi: 10.1111/cogs.12265

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends Cogn. Sci.* 23, 408–422. doi: 10.1016/j.tics.2019.02.006

Botvinick, M., and Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.* 16, 485–488. doi: 10.1016/j.tics.2012.08.006

Campbell, J. O. (2016). Universal darwinism as a process of bayesian inference. *Front. Syst. Neurosci.* 10, 49. doi: 10.3389/fnsys.2016.00049

Candadai, M., and Izquierdo, E. J. (2020). Sources of predictive information in dynamical neural networks. *Sci. Rep.* 10, 1. doi: 10.1038/s41598-020-73380-x

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philos. Transact. R. Soc. B Biol. Sci.* 362, 1585–1599. doi: 10.1098/rstb.2007.2054

Clark, A., and Chalmers, D. J. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Constant, A., Clark, A., and Friston, K. J. (2020). Representation wars: enacting an armistice through active inference. *Front. Psychol.* 11, 598733. doi: 10.3389/fpsyg.2020.598733

Crouse, M., Nakos, C., Abdelaziz, I., and Forbus, K. (2020). *Neural Analogical Matching.* arXiv:2004.03573 [Cs]. Available online at: http://arxiv.org/abs/2004.03573 (accessed September 24, 2023).

Damasio, A. (2012). *Self Comes to Mind: Constructing the Conscious Brain (Reprint Edition)*. New York, NY: Vintage.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philos. Psychol.* 20, 247–268. doi: 10.1080/09515080701239510

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., et al. (2023). *PaLM-E: An Embodied Multimodal Language Model (arXiv:2303.03378)*. arXiv.

Dumont, N. S.-Y., Furlong, P. M., Orchard, J., and Eliasmith, C. (2023). Exploiting semantic information in a spiking neural SLAM system. *Front. Neurosci.* 17, 1190515. doi: 10.3389/fnins.2023.1190515

Friston, K., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M., et al. (2023). Path integrals, particular kinds, and strange things. *Phys. Life Rev.* 47, 35–62. doi: 10.1016/j.plrev.2023.08.016

Gentner, D. (2010). Bootstrapping the mind: analogical processes and symbol systems. *Cogn. Sci.*, 34, 752–775. doi: 10.1111/j.1551-6709.2010.01114.x

George, D., Rikhye, R. V., Gothoskar, N., Guntupalli, J. S., Dedieu, A., and Lázaro-Gredilla, M. (2021). Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. Commun.* 12, 1. doi: 10.1038/s41467-021-22559-5

Hawkins, J. (2021). *A Thousand Brains: A New Theory of Intelligence*. New York, NY: Basic Books.

Heylighen, F. (2016). Stigmergy as a universal coordination mechanism I: definition and components. *Cogn. Syst. Res.* 38, 4–13. doi: 10.1016/j.cogsys.2015.12.002

Hipólito, I., Winkle, K., and Lie, M. (2023). Enactive artificial intelligence: subverting gender norms in human-robot interaction. *Front. Neurorobot.* 17, 1149303. doi: 10.3389/fnbot.2023.1149303

Hofstadter, D. R. (2007). *I Am a Strange Loop*. New York, NY: Basic Books.

Hutto, D., and Hipólito, I. (2021). Updating our theories of perceiving: from predictive processing to radical enactivism. *PsyArXiv*. doi: 10.31234/osf.io/x3nka

Ji, X., Elmoznino, E., Deane, G., Constant, A., Dumas, G., Lajoie, G., et al. (2023). *Sources of Richness and Ineffability for Phenomenally Conscious States (arXiv:2302.06403)*. arXiv.

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.* 112, 323–343. doi: 10.1007/s00422-018-0753-2

Kolchinsky, A., and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* 8, 20180041. doi: 10.1098/rsfs.2018.0041

Kosiorek, A., Sabour, S., Teh, Y. W., and Hinton, G. E. (2019). "Stacked capsule autoencoders," in *Advances in Neural Information Processing Systems*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.), 15512–15522. Available online at: http://papers.nips.cc/paper/9684-stacked-capsule-autoencoders.pdf (accessed September 24, 2023).

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. J. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Front. Robot. AI* 5, 21. doi: 10.3389/frobt.2018.00021

Oliver, M. (1986). "*Poem (the spirit likes to dress up)" from Dream Work*. Boston, MA: Atlantic Monthly Press.

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. *A Bradford Book* Cambridge, MA: MIT Press.

Ramstead, M. J. D., Badcock, P. B., and Friston, K. J. (2018). Answering Schrödinger's question: A Free-energy formulation. *Phys. Life Rev.* 24, 1–16. doi: 10.1016/j.plrev.2017.09.001

Ramstead, M. J. D., Kirchhoff, M. D., and Friston, K. J. (2019). *A Tale of Two Densities: Active Inference is Enactive Inference*. Available online at: http://philsci-archive.pitt.edu/16167/ (accessed September 24, 2023).

Safron, A. (2019). *Multilevel Evolutionary Developmental Optimization (MEDO): A Theoretical Framework for Understanding Preferences and Selection Dynamics. arXiv:1910.13443 [Econ, q-Bio, q-Fin]*. Available online at: http://arxiv.org/abs/1910.13443 (accessed September 24, 2023).

Safron, A. (2021a). The radically embodied conscious cybernetic bayesian brain: from free energy to free will and back again. *Entropy* 23, 6. doi: 10.3390/e23060783

Safron, A. (2021b). World models and the physical substrates of consciousness: hidden sources of the stream of experience? *J. Conscious. Stud.* 28, 210–221. doi: 10.53765/20512201.28.11.210

Safron, A. (2022). Integrated world modeling theory expanded: implications for the future of consciousness. *Front. Comput. Neurosci.* 16:642397. doi: 10.3389/fncom.2022.642397

Safron, A., Çatal, O., and Verbelen, T. (2021). *Generalized Simultaneous Localization and Mapping (G-SLAM) as Unification Framework for Natural and Artificial Intelligences: Towards Reverse Engineering the Hippocampal/Entorhinal System and Principles of High-Level Cognition*. PsyArXiv.

Silver, D., Singh, S., Precup, D., and Sutton, R. S. (2021). Reward is enough. *Artif. Intell.* 299, 103535. doi: 10.1016/j.artint.2021.103535

Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness*. New York, NY: W. W. Norton & Company.

Tani, J. (2016). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. Oxford: Oxford University Press.

Thomas, V., Bengio, E., Fedus, W., Pondard, J., Beaudoin, P., Larochelle, H., et al. (2018). *Disentangling the Independently Controllable Factors of Variation by Interacting With the World*. arXiv.

Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., et al. (2017). *Independently Controllable Factors*. *arXiv*.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

van Gelder, T. (1990). *Why Distributed Representation is Inherently Non-symbolic*. Available online at: https://philpapers.org/rec/VANWDR (accessed September 24, 2023).

Van Gelder, T. (1995). What might cognition be, if not computation? *J. Philos.* 92, 345–381. doi: 10.2307/2941061

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., and Kirmayer, L. J. (2019). Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 43, 1–97. doi: 10.1017/S0140525X19001213

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Curr. Opin. Behav. Sci.* 38, 90–95. doi: 10.1016/j.cobeha.2021.01.002

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., et al. (2020). The Tolman-Eichenbaum Machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* 183, 1249–1263.e23. doi: 10.1016/j.cell.2020.10.024

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., et al. (2023). *A Survey on Multimodal Large Language Models (arXiv:2306.13549)*. arXiv.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 05 October 2021
doi: 10.3389/fnbot.2021.729665

# The Acquisition of Culturally Patterned Attention Styles Under Active Inference

*Axel Constant*[1]*, *Alexander Daniel Dunsmoir Tschantz*[2,3], *Beren Millidge*[2,4], *Felipe Criado-Boado*[5], *Luis M Martinez*[6], *Johannes Müeller*[7] and *Andy Clark*[2,8,9]

[1] Theory and Method in Biosciences, University of Sydney, Sydney, NSW, Australia, [2] Department of Informatics, The University of Sussex, Sussex, United Kingdom, [3] Sackler Centre for Consciousness Science, University of Sussex, Brighton, United Kingdom, [4] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [5] Institute of Heritage Sciences, Spanish National Research Council, Santiago de Compostela, Spain, [6] Institute of Neurosciences, Spanish National Research Council, Universidad Miguel Hernández, Alicante, Spain, [7] Institute of Prehistoric and Protoshistoric Archaeology, Kiel University, Kiel, Germany, [8] Department of Philosophy, The University of Sussex, Sussex, United Kingdom, [9] Department of Philosophy, Macquarie University, Sydney, NSW, Australia

This paper presents an active inference based simulation study of visual foraging. The goal of the simulation is to show the effect of the acquisition of culturally patterned attention styles on cognitive task performance, under active inference. We show how cultural artefacts like antique vase decorations drive cognitive functions such as perception, action and learning, as well as task performance in a simple visual discrimination task. We thus describe a new active inference based research pipeline that future work may employ to inquire on deep guiding principles determining the manner in which material culture drives human thought, by building and rebuilding our patterns of attention.

**Keywords: perception, culture, archaeology, active inference, eye tracking, simulation**

OPEN ACCESS

***Edited by:***
*Jun Tani,*
*Okinawa Institute of Science and Technology Graduate University, Japan*

***Reviewed by:***
*Alejandra Ciria,*
*Universidad Nacional Autónoma de México, Mexico*
*Shingo Murata,*
*Keio University, Japan*

***Correspondence:**
*Axel Constant*
*axel.constant.pruvost@gmail.com*

***Received:*** *23 June 2021*
***Accepted:*** *02 September 2021*
***Published:*** *05 October 2021*

***Citation:***
*Constant A, Tschantz ADD, Millidge B, Criado-Boado F, Martinez LM, Müeller J and Clark A (2021) The Acquisition of Culturally Patterned Attention Styles Under Active Inference. Front. Neurorobot. 15:729665. doi: 10.3389/fnbot.2021.729665*

## INTRODUCTION

Do the worlds we build alter our own minds and the ways we process information? In one sense, it is obvious that they do—we read books, we listen to our teachers, and learn new ways of thinking and reasoning as a result. Thanks to lifelong learning, we may become experts in a domain such as forestry and become able to attend to, and differentiate, new things as a result. But we are also immersed in a sea of material structures and artefacts such as pottery, ceramics, clothing, buildings, tools, and more. As we encounter and explore these artefacts and structures, they too influence our patterns of visual and embodied exploration, and thus our learning. But the nature and potential cognitive importance of these interactions with material structure remains ill-understood.

Iterated encounters with non-linguistic aspects of material culture, we believe, do not simply reflect human thinking and reasoning—rather, they shape and alter it. Our minds are as much the products of these materialities as the cause. This is a bold claim, yet one that is quite often found in the sciences of mind and culture (Dennett, 1991, 1996; Clark, 1997; Sutton, 2002; Knappett, 2005; Renfrew and Malafouris, 2010). To our knowledge, it is a claim that has not been experimentally demonstrated or subjected to rigorous analysis and testing. The simulations we report below are meant as a first step towards building a pipeline to explore and test this claim—that encounters with non-linguistic artefacts can alter patterns of thought and attention in cognitively interesting and beneficial ways.

With this goal in mind, our paper presents a proof of principle for modelling visual foraging and sensory learning of artefacts using active inference for Markovian inference models. Markovian models are used to perform predictive statistical inference over some states of interest, given the outcomes those states are known to generate. For instance, Markovian model can be used

to perform weather forecasting over some unknown states (a.k.a. hidden states, or unknown variables) which would represent the weather for each day of the week (e.g., rainy; sunny), and where the outcomes would be some observable property of the possible states (e.g., cloud shapes). After observing these properties, inference proceeds by combining known prior probabilities of transitioning from one state to the next (e.g., history of transitions between rainy to sunny days, or sunny days to sunny days, etc.) with the known likelihood of observables under each state. The resulting posterior specifies the Bayesian probability the hidden variables at the next time step (e.g., tomorrow's weather). In this paper, we utilise the theory of active inference to perform the requisite inference, action-selection, and learning for our model.

Crucially, active inference straddles multiple levels of Marr's hierarchy, from deep computational considerations in statistical inference and thermodynamics, all the way down to being able to build biologically plausible models of psychophysical tasks which accord with known neurophysiology. This theoretical breadth licences us to create an active inference model of the interaction between inference and learning systems and material culture which can demonstrate, through psychophysical observables, the importance of the material world for shaping mental attention styles and ultimately cognitive capacities.

We present two experiments. In experiment 1, we provide a pilot simulation of visual foraging that showcases the potential of our method for modelling empirical data on foraging over differentially complex cultural artefacts. Experiment 1 can be viewed as a training experiment, where artificial agents learn about the hierarchical structure of artefacts, and where this acquired knowledge is later applied to a categorisation task (experiment 2). Our simulation focuses on implementing artificial behaviour that could mimic *in vivo* participants scan paths over antique vases such as observed in Criado-Boado et al. (2019). Criado-Boado and colleagues studied the influence of decoration patterns on scan paths employed by visual foragers. They use a vertical index (Vi) to measure the influence of various patterns on visual saccades, relative to the size of the visual display presenting the differentially decorated vases (e.g., more or less complex decoration painted horizontally or vertically).

We show that an increase in decoration complexity, when modelled as patterns of hidden states, entails characteristically different scan paths, and hence Vi; We call these scanpaths "Culturally Patterned Attention STyles" (C-PAST). These scan paths are the result of the agent attempting to predict the next decoration based on observed pigments and learning the probability transitions between the visual motifs forming the decorations. These scanpaths should be viewed as heuristics of culturally shaped patterns of attention. Future work should attempt to fit the model generating those scanpaths with real participant data. The motivation for calling the scanpaths "cultural" is that vase decorations have been shown to be a good indicator of cultural complexity. Hence, we call "cultural," or rather "culturally patterned" the scanpaths that result from learning based on the exposure to such decorations; the patterning here being synonym of learning.

In experiment 2, we provide a modelling method to transfer the learning of priors across simulations under active inference, which, to our knowledge, has never been done before in the literature. Transfer learning here refers simply to the transfer of knowledge across tasks (e.g., employing knowledge acquired in task X to perform the actions required in task Y). The challenge with the transfer of learning in active inference modelling is that the model parameters (e.g., transition probabilities between hidden states) are normally task specific, which means that they correspond to the environment of the task at hand (e.g., motifs as hidden states being specific to the vase perceived by the agent). The novel modelling strategy we propose in this paper allows transfer learning by breaking down the environment of a task into units that are general purpose hidden states. These units are locations in a discrete 2-dimensional map, which we call the remapping likelihood matrix (see method for details). The remapping matrix allows us to local representations of the immediate environment, and, crucially, to reuse these units or groups of units, when learned, across tasks. While it unlocks the possibility to accomplish our simulation, we recognise that the present method of likelihood remapping is trivial. Based on the learning of the structure of vases' decorations in experiment 1, in experiment 2 we simulate a pattern categorisation task that involves reusing learned model parameters in experiment 1. In the categorisation task, the agent has to match a series of motif cut-outs with their corresponding motif. We show how performance (hits vs. non-hits) changes depending on learned parameters under the different levels of cultural complexity afforded by vases transferred from experiment 1.

In summary, with experiment 1 and 2, we show the potential of active inference to study (i) exposure to artefactual complexity leading to the acquisition of the knowledge underwriting different Culturally Patterned Attention STyles (C-PAST)—here knowledge about transition probabilities among hidden states, or representations of the structure of the world; (ii) the repurposing of C-PAST in novel cognitive tasks, and the manner in which different C-PAST influence performance in novel cognitive tasks. We are aware that the task that we use may be considered too simple to demonstrate the effect of C-PAST on cognitive task performance, and that our task is limited to non-natural scenes. However, the goal of our simulation, beyond reproducing the results of Criado-Boado et al. (2019) is to provide a simple example of a scalable modelling strategy for future research on related issues in the field of cognitive archaeology.

## VERTICAL INDEX, SOCIAL COMPLEXITY, CULTURAL COMPLEXITY AND ATTENTION

The vertical index (Vi) is a measure that compares the proportion of horizontal to vertical saccades made when viewing an image (Criado-Boado et al., 2019; Millán-Pascual et al., 2021). This measure, which is closely related to the density of information presented in vertical dimensions, has been shown to vary considerably across items ranging from pots to monuments, drawn from different archaeological epochs (Prieto-Martínez

et al., 2003). The archaeological record shows that decoration patterns of complex prehistoric societies generally followed high Vi patterns, whereas low Vi patterns are found in simpler societies. Criado-Boado et al. (2019) note that archaeologists accept that the evolution of pottery decorations parallels, in the particular chronological sequence on study, changes in the level of complexity of social organisation [see Prieto-Martínez et al. (2003) for a detailed characterisation of the social complexity embedded in the pottery sequence analysed in Criado-Boado et al. (2019) and see Criado-Boado (2014)] for a more general and theoretical account of the interactions between materiality and social processes] and propose that a virtual index of decoration may be a measure or reflection of such a social complexity [Müller et al. (2015) also illustrates similar conclusions for a different pottery style]. Criado-Boado et al. (2019) found that the verticality of decoration correlated with the chronological evolution of the decorations on ceramics displayed in their study; these being associated with difference cultural periods and associated levels of social complexity. They showed that eye movements of participants followed the same evolutionary trend reported by the Vi index when presented with the vases' decorations characteristic of each successive social periods.

Here, what we refer to as social complexity differs from what is sometimes described as cultural complexity (Sterelny, 2020). Social complexity denotes the overall level of organisation of a society, whereas cultural complexity denotes the complexity of artefacts found in a given population. Cultural complexity is sometimes viewed as a proxy to social complexity, as it would reflect the level of skills and expertise of the tool and artefacts makers and users, which in turn would reflect the level of social complexity. Cultural complexity can be viewed as either repertoire complexity, or peak complexity (Sterelny, 2020). Repertoire complexity corresponds to the number of distinct tools that were used in each society, whereas peak complexity corresponds to the level of complexity of a given tool, which can be measured in terms of parts and functions of the tool; these being called technounits Oswalt (1973). The correlation observed by Criado-Boado et al. (2019) was between social complexity and the Vi of decorations on vases. The correlation was not between social complexity and cultural complexity.

A challenge with studying the relation between social complexity and cultural complexity is that repertoire complexity and peak complexity may vary independently (Sterelny, 2020), and depending on the sort of artefact one considers, peak complexity may even be inversely proportional to the true level of skills of artefacts makers reporting social complexity. Moreover, the locus of peak complexity may change over time in a same society. These problems are especially salient when considering the complexity of aesthetic objects like vase patterns. For instance, it is common to observe disparity within the artefactual repertoire, with simpler societies having poorer decorative vase patterns but highly complex body ornamentations like tattoo motifs or plumes arrangements. The same applies to more advanced societies and pottery decorations, whose peak complexity can correlate at first with the level of social complexity, but then decrease with time as the society discovers new material and media for artistic expression (e.g., jewellery, metallurgy, architecture, etc.). For instance, pottery was important to express social styles and social identities in the Atlantic façade between 6,000 and 2,000 BP, while in other cultures and times other sort of material were used to mainly express social identity (e.g., jewellery, metallurgy, monuments, or tattoos, personal ornaments, or plumes).

Despite the intrinsic interest of these issues and their importance for understanding the historical record, it is important to note that our target in the simulation studies is something rather different. Our goal is to explore the potential role of cognition (attention, perception and learning) as a variable operating within these complex regimes. Specifically, we are asking whether, and in what ways, interactions with artefacts might alter patterns of attending, which in turn alter ways of thinking and reasoning about the world (e.g., in a cognitive task). Thus, we introduce cognition (attention, perception and learning) as a third variable to the complex relation between social complexity and cultural complexity. The hope is that styles of cognition may function as an explanatory bridge between cultural and social complexity. Accordingly, our simulation explores the synthetic relationship between task performance and the acquisition or learning of attention styles based on the exposure to vase decorations. The motivation for this simulation is to explore the ways interactions with artefacts might alter patterns of attending, which in turn alter ways of thinking and reasoning about the world (e.g., in a cognitive task). If such effects are real, then there may be a good reason to believe that there exists a link between the structure of the human-made world and the ways we think and reason after cultural immersion in different such worlds. Because artefacts affording greater vertical index correlate with social complexity, and because vertical indices illicit characteristically different visual foraging patterns (Criado-Boado et al., 2019), one could hypothesise that there is a ratchetting loop between the acquisition of attention styles, features of the artefacts that illicit such an acquisition (e.g., Vi), and cultural complexity.

Note that novel patterns of attending do not necessarily witness of a neurobiological change in the human evolutionary history (e.g., encephalization). The hypothesis on the cognition-culture loop is not primarily a gene-culture co-evolutionary hypothesis on the evolution of social complexity and cultural complexity (e.g., Henrich, 2015). Rather, such a hypothesis refers to dynamics at the level of cognition and culture. Attention styles are acquired over developments; they are akin to cognitive "gadgets" (Heyes and Frith, 2014) that support the scaf-folding of more complex abilities such as language and mind reading. One of these abilities may be that of reproducing complex human social ensembles; an ability scaf-folded through artefactually mediated acquisition of attention styles. The current simulation is a first step towards studying such cognition-culture loop under the theory of active inference.

Finally, note that our project differs from related research in the field of active inference, culture and cognition. Here, our goal is not to account for the formation and function of human culture, but rather, to inquire on the manner in which culture shapes perception and influences task performance. That is, we are not here attempting to define what culture is and

how it works, but rather we are here attempting to describe the way humans may respond to its products and how those influence cognitive task performance. While this latter problem is certainly part of the more general project of accounting for the formation and function of culture, this problem remains one that can be approached independently of the larger discussion on the ontology of culture. Under active inference, the ontology of culture is defined as patterns of attention, or "regimes" of attention shaped by local practises (e.g., Kaufmann and Clément, 2007; Ramstead et al., 2016; Constant et al., 2019, 2020; Veissière et al., 2020). Despite the differences in research orientations noted above, our simulation may be viewed as providing one possible illustration of the manner in which the acquisition of regimes of attention (here C-PAST) influences task performance.

## METHOD: ACTIVE INFERENCE

Active inference is a theory arising from theoretical neuroscience, which posits that perception, action, and learning can be fundamentally united since they can be cast as performing a form of approximate Bayesian inference (known as variational inference) on the same information—theoretic objective (Friston, 2010). Although anchored in abstract conceptions of inference, active inference possesses a neurobiologically plausible process theory (Friston et al., 2017), and has been applied to explaining and building models of diverse aspects of neural and cognitive function such as planning and navigation under uncertainty (Kaplan and Friston, 2018), saccade generation and reading (Parr and Friston, 2017), sequential decision making tasks (Friston et al., 2013, 2016), up to complex continuous control tasks (Pio-Lopez et al., 2016; Fountas et al., 2020; Millidge, 2020; Tschantz et al., 2020), as well as psychophysical observables such as modelling evidence accumulation (FitzGerald et al., 2015). Moreover, through the expected free energy functional, active inference also entails a natural epistemic drive which has been exploited before in previous active-inference studies of visual foraging (Friston et al., 2015; Mirza et al., 2016). Here, we present a high-level description of active inference. For a detailed overview of active inference in discrete state-spaces and for the purpose of economy of space, we refer the technically minded reader to the dedicated method papers of Friston et al. (2015, 2017), and Da Costa et al. (2020).

### An Overview of Variational Inference

Active inference posits that action, learning and perception can all be described as a process of variational inference. Variational inference is an approximation to exact Bayesian inference which postulates the existence of a variational recognition density, which is matched to the true posterior via an optimisation process. Variational inference thus converts a difficult and intractable inference procedure into a potentially tractable optimisation process, for which good approximate solutions exist. Variational inference obtains its solution by minimising the variational free energy functional, and this is used in our model for perception—i.e., the inference of hidden states from observed outcomes. Active inference extends this theory to include action, which is inferred from preferences over sequences of potential future states. This requires the use of a subtly different objective functional—the expected free energy—which is a functional over expected future states and observations. The expected free energy naturally includes an epistemic exploration-inducing information gain term which encourages active inference agent's to seek out novel outcomes, which thus can mimic key behaviours in visual foraging which is all about information gathering.

Variational inference depends on two mathematical objects—the variational recognition distribution (hereafter referred to as the variational distribution) and the generative model. The variational distribution is a distribution over all hidden variables in the model and represents the agent's beliefs about the state of the world. The generative model is the agent's model of how the observables in the world are "generated" by the hidden variables which must be inferred. During inference, the variational distribution (the agent's beliefs) are optimised to best conform to the outcomes or data observable by the agent. Thus, in inference, the generative model is "inverted" —in that we look to recover the mapping from observations to hidden states, given a mapping from hidden states to observations.

### An Overview of the Variational Distribution

Formally, let x refer to hidden variables, where $x^{n}$ refers to the hidden state at level n, and $\pi$ refer to a policy (fixed sequence of actions). The variational distribution can then be factorised as follows:

$$Q(x^1_{0:T}, x^2_{0:T}, \pi) = Q(x^1_0)Q(x^2_0)Q(\pi)\prod_{t=1}^{T} Q(x^1_t|\pi)Q(x^2_t|\pi) \quad (1)$$

Moreover, the agent's generative model can be factorised as:

$$p(o_{0:T}, x^1_{0:T}, x^2_{0:T}, \pi) = p(\pi)p(x^1_0)p(x^2_0)p(o_0)$$
$$\prod_{t=1}^{T} p(o_t|x^1_t)p(x^1_t|x^1_{t-1}, x^2_t, \pi)p(x^2_t|x^2_{t-1}, \pi) \quad (2)$$

Given these distributions, inference is achieved by optimising the variational distribution in order to minimise free energy:

$$Q*(x^1_{0:T}, x^2_{0:T}) = \text{argmin}F(Q_{0:T}, o_{0:T})$$
$$F = D_{KL}(Q(x^1_{0:T}, x^2_{0:T}; \phi)||p(o_{0:T}, x^1_{0:T}, x^2_{0:T})) \quad (3)$$

In a similar fashion, action selection is achieved by optimising the variational distribution to minimise expected free energy, which we compute at each step:

$$\pi \sim Q^*(\pi) = \arg\min_{\pi} \varsigma(Q, \pi)$$
$$\varsigma(Q, \pi) = E_{Q(o_{t:T}, x^1_{t:T}, x^2_{t:T}|\pi)}\left[D_{KL}Q(x^1_{t:T}, x^2_{t:T}|\pi) \right.$$
$$\left. ||p(o_{t:T}, x^1_{t:T}, x^2_{t:T})\right]$$

### An Overview of the Generative Model

In the current work, the active inference agent utilises a two-level hierarchical generative model parametrised by four matrices "$A^1$," "$A^2$" and "$B^{1''}$ and "$B^{2''}$ (for a deeper description of hierarchical models in active inference see, Friston et al., 2017). Here we present the role these matrices play in the variational inference over states, future states and outcomes in general. In the result section, we describe the semantic of these matrices, which

will be specific to the tasks we seek to accomplish in experiments 1 and 2.

The "A" matrices represent the parameters of a likelihood distribution which maps from the hidden states at a hierarchical layer to the outcomes associated with the layer (the outcomes of all hierarchical layers other than the lowest layer correspond to the hidden states of the hierarchical layer below). These matrices denote the instantaneous probabilistic mappings between the hidden states and outcomes.

The "B" matrices represent the (policy- dependent) transitions between the hidden states over multiple time-steps. The parameters of the "B" matrices were learnt through experience. This learning can be cast as inference on the parameters of a dirichlet hyperparameter over the entries of the "B" matrix. For more details see Da Costa et al. (2020). Crucially, the parameters of the "$B^{2''}$ matrix are inferred over the course of a trial and provide the representation of the C-PAST.

Given such a generative model and an initial state distribution, sequences of potential future outcomes and hidden states can be generated and compared for different potential policies (sequences of action) which could be enacted. These sequences of future outcomes and hidden states are scored by the expected free energy functional (denoted "G"). Policies are selected which minimise "G.". In our experiments, look-ahead was only performed for a single time step into the future and actions were selected which greedily minimised "G."

Another important aspect is the encoding of an agent's preferences into the generative model. This is encoded through the matrix "C" which specifies a desired probability distribution over outcomes. In our experiments the agent strongly desired to observe pigments and will be averse to observing non-pigments. This simple constraint on agent behaviour is sufficient to generate complex visual-foraging behaviour.

We specify a prior the entries of the "A" matrices. Crucially, to ensure that the active inference agent only was in possession of local knowledge (i.e., the content of its foveated region and not the entire image), we utilised the novel likelihood remapping trick by which the "A" matrices were represented in a state-dependent fashion so that the agent was only aware for a given hidden state (location), the presence or absence of the pigment in a 3 x 3 square around the agents location. Likelihood remapping allows the agent to perform state inference and navigation by bypassing the full representation of the generative process (i.e., environment). This is in contrast to standard active inference approaches which typically require the agent to be given a correct global understanding of the scene. To achieve this locality, the "A" matrix becomes state-dependent so that it only provides information about outcomes in the proximity of the state the agent is in. A further description of this likelihood remapping method can be seen below in **Figure 1**.

# RESULTS

## Experiment 1

The goal of experiment 1 was to exemplify the relation between artefactual complexity and scan path cultural specificity under active inference. We showed how variations in artefactual complexity leads to the acquisition of different "Culturally Patterned Attention STyles" (C-PAST). Scan paths are artificial visual saccades enacted by the agent during the visual foraging task. The goal of the visual foraging task was simply for the agent to explore the visual scene, which consists of a vase decorated with motifs made of pigments. The agent's simulated gaze starts at the centre of the vase and is free to explore the vase for 100 timesteps. We presented the simulated agent with vases that had different levels of complexity—that is, that were made up of more or less visually rich patterns. The richness of the patterns came from the inclusion of more or less vertical features, or motifs, from horizontal (0 degree angle), to oblique (−45 and +45 degree angle), to vertical (90 degree angle). We measured the influence of pattern complexity on visual foraging with a version of the virtual index (Vi) used in Criado-Boado et al. (2019). Vi is a measure of visual saccades relative to the size of the display upon which the vase is presented. The empirical results of Criado-Boado et al. (2019) suggest that vase complexity affects change in scan paths' Vi. The purpose of simulation 1 was to reproduce this effect in silico and based on the parameters needed to simulate the effect, phenotype the different attention style or C-PAST acquired through exposure to vases with four levels of complexity (0 to 3) (see **Figure 2**).

### The Model for Experiment 1

To perform the task in experiment 1, the simulated agent applies our inference algorithm to a simple two level Markovian generative model. The generative model allows the agent to infer two things: (i) the hidden states at level 1 or 2, and (ii) an action policy, which optimises the desired sequence of hidden states enacted by the agent. In experiment 1, the level 1 hidden states are locations on the visual display where pigments can be found. The presence or absence of a pigment functions as the sensory outcome. Level 2 hidden states represent the motifs which consist of repeating patterns of pigments, for instance, crosses, diagonal and horizontal lines (see **Figure 2**). Using a two-level hierarchical generative model allows us to simulate an agent that can infer the presence of more abstract hidden states (i.e., level 2 motifs) based on its inference of simpler hidden states (i.e., level 1 pigments). For each cycle of inference at level 2, four cycles of inference are performed at level 1, that is, four pigments are inferred. The heatmap we present below is the result of having inferred those different hidden states. The second thing the agent can infer based on its generative model is an action policy, which here stands for a (sequence of) visual saccades. Action policies are simply sequences of control states that are inferred over multiple time steps based on preferences the agent has for certain outcomes.

The generative model represents and performs inference over four sets of parameters. The first is a likelihood parameter $A^1$, which exists at level 1, and is a probabilistic mapping between sensory outcomes (pigments) and level 1 hidden states (locations on the visual display). At level 1, we keep the likelihood deterministic (all [0 1]), which speaks to the fact that the agent can clearly perceive the pigments. The second likelihood parameter $A^2$, represents the probabilistic mapping between the motifs and the locations perceivable by the agent. This likelihood

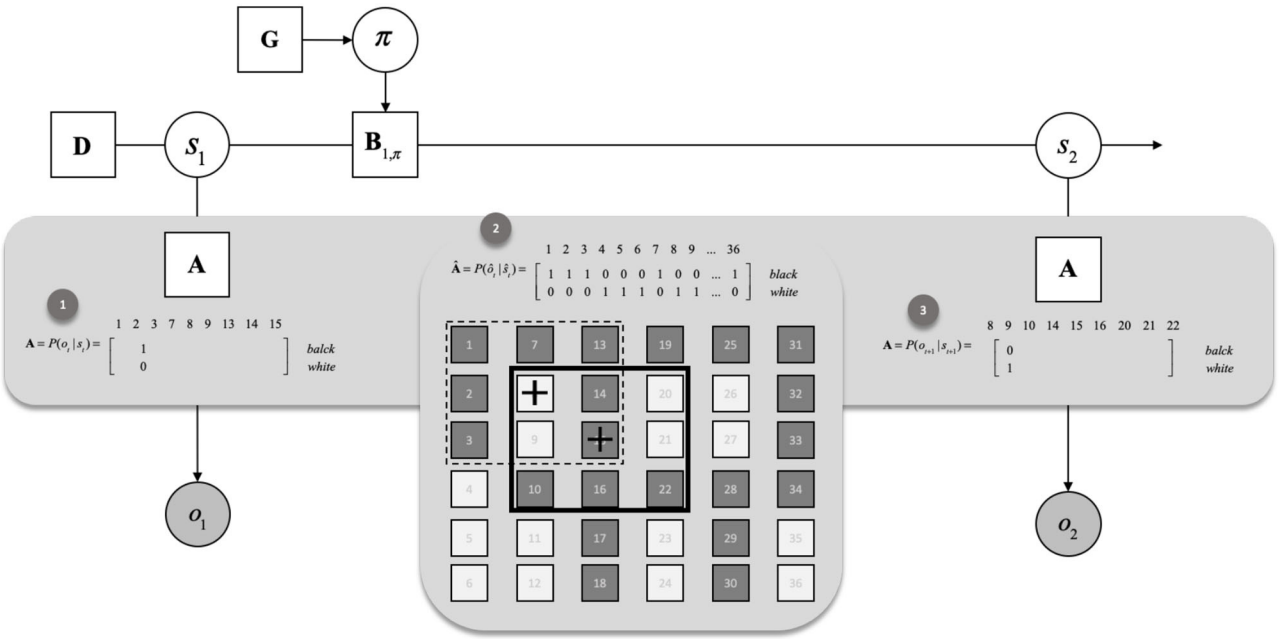



**FIGURE 1 |** Likelihood remapping. For illustrative purposes, this figure presents a generative process (A hat) made of 36 locations, or states. Each state is associated with an outcome, either black or white. (1) The initial likelihood is defined for an agent that would start in location 8. The likelihood of the generative model is specified based on the 8 locations, or hidden states surrounding the current location, as well as the current location. (2) Based on the inferred policy (e.g., 8 → 15), we move the agent in the generative process, here, to location 15. (3) Indexing the novel surrounding and current location from the generative process, we remap the likelihood that will be used at t + 1 to infer the state and the policy.

is also deterministic, which speaks to the fact that the agent knows how a given motif (for instance a cross) can be represented by a sequence of pigments. The second set of parameters are the transition probability mappings between level 1 hidden states $B^1$ or the level 2 hidden states $B^2$. In experiment 1, the agent learns the transitions between hidden states (motifs) at level 2. Transitions are deterministic at level 1 and depend entirely on the action policy. A two-level model, with uncertainty in level 2 transitions (motifs transitions) will scan differently. Increases in vase complexity drive the learning of motifs transition (i.e., patterns). We refer the reader to the method section for the details of the manner in which our inference algorithm performs the inferences, formulates action policies, and learns $B^2$.

## Vertical Index

The Vertical index (Vi) is defined as the height "h" of the region upon which the agent gazed times the number of vertical saccades (number of steps taken vertically given the inferred policies), minus the width "w" of the region gazed upon time the number of horizontal saccades, all that divided by the sum of the product of the height "h" and number of vertical saccades, and the product of the width "w" and the number of vertical saccades:

$$VirticalIndex(Vi) = \frac{h{*}nb.vert.saccades - w{*}nb.horiz.saccades}{h{*}nb.vert.saccades + w{*}nb.vert.saccades} \quad (4)$$

The change in Vi relative to the four levels of complexity are presented in **Figure 3**, with their associated level of decoration complexity. GIF representations of the simulation as well as the source code for all experiments can be found at https://github.com/BerenMillidge/MaterialCulture. The results show that Vi correlates positively with the levels of complexity, as expected, and empirically observed in Criado-Boado et al. (2019). We present the scanpaths in **Figure 4**.

## C-PAST

We define the "Culturally Patterned Attention STyles" (C-PAST) as sets of motif-transition parameters learned when the agent is presented with decorations during exploration. Our simulation shows that difference in pattern complexity naturally entails differences in C-PAST, leading to systematic differences in Vi (**Figure 3**). To measure the C-PASTs, we use the entropy, in information theoretic terms, of the sets of motif-transition parameters B2. Formally, we define our measure of C-PAST as

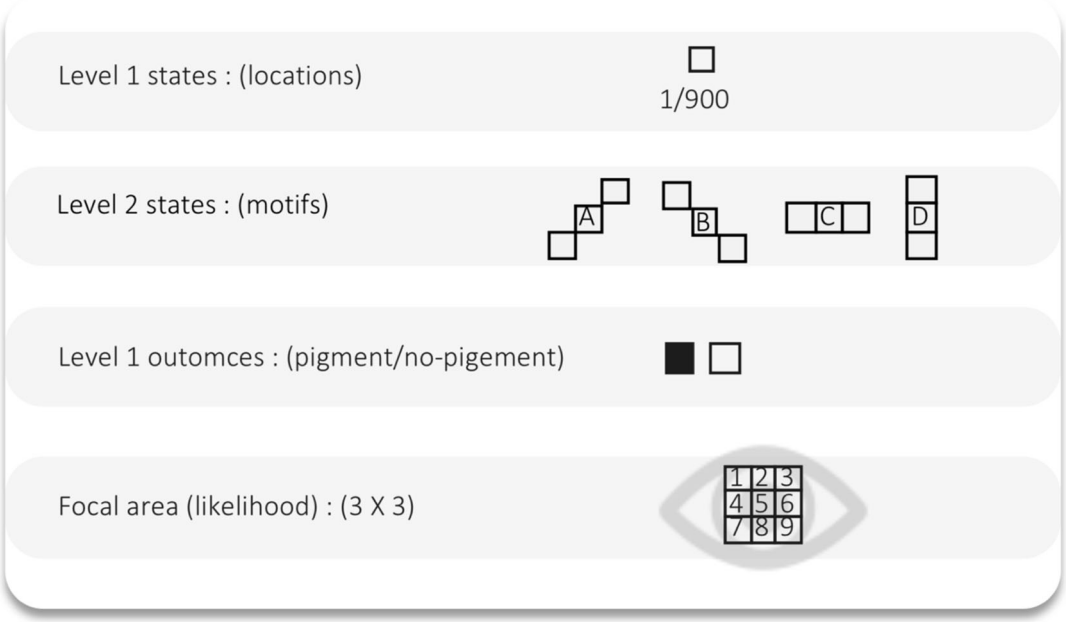
$$C - PAST = H[B^2] = -\sum_{i}^{N} B^2_{:i} \log B^2_{:i} \quad (5)$$

where H is the Shannon entropy and N is the number of motifs. We use entropy because it allows us to describe intrinsic features of the distributions, without having to commit to a normative assessment of those distributions (e.g., compared to an ideal, extrinsic criterion of goodness). Indeed, the purpose of measuring C-PASTs is simply to phenotype the various attention styles that obtain from the exposure to various levels of cultural complexity. Note that we only allowed for learning of transition probabilities (B2 parameters), but in principle, nothing prevents

A

$$VirticalIndex(Vi) = \frac{h * nb.vert.saccades - w * nb.horiz.saccades}{h * nb.vert.saccades + w * nb.vert.saccades}$$

B




C

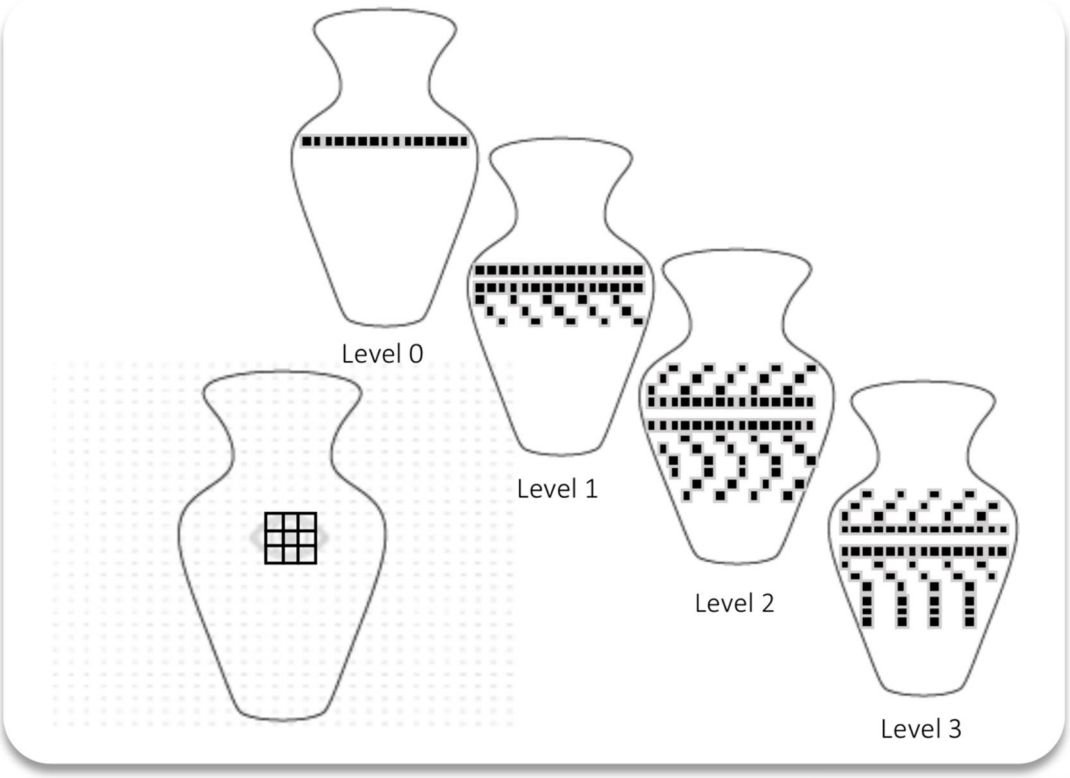



**FIGURE 2 | (A)** Vi is a measure of visual saccades relative to the size of the visual area gazed upon. This area increases with decoration complexity. **(B)** States and outcomes for the generative model. At level 1, the outcomes are the absence or presence of a pigment. The states are the locations (1/900) of which we

*(Continued)*

**FIGURE 2 |** take a sample of nine currently available locations to define the focal area (updated based on the navigation matrix for every trial, see method). Level 2 states are motifs (sets of locations) and their associated outcomes are the location at level 1 [e.g., "P(location|motif)]. The focal area corresponds to the outcome likelihood at level 1. The important thing to remember is that we are respecifying the likelihood after each eye movement, using the remapping likelihood matrix (see method section). **(C)** 900 locations grid over which the active inference agent scans. The agent can decide to move from the central location of the 3 x 3 focal area (grid) to any location of the grid. The four levels of decoration complexity build on one another. Level 0 is a straight line, and level 1 adds verticality by adding oblique shapes below the line. Level 2 adds oblic motifs on the top of the straight line as well, and level 4 adds vertical lines below the oblique shapes at the bottom.

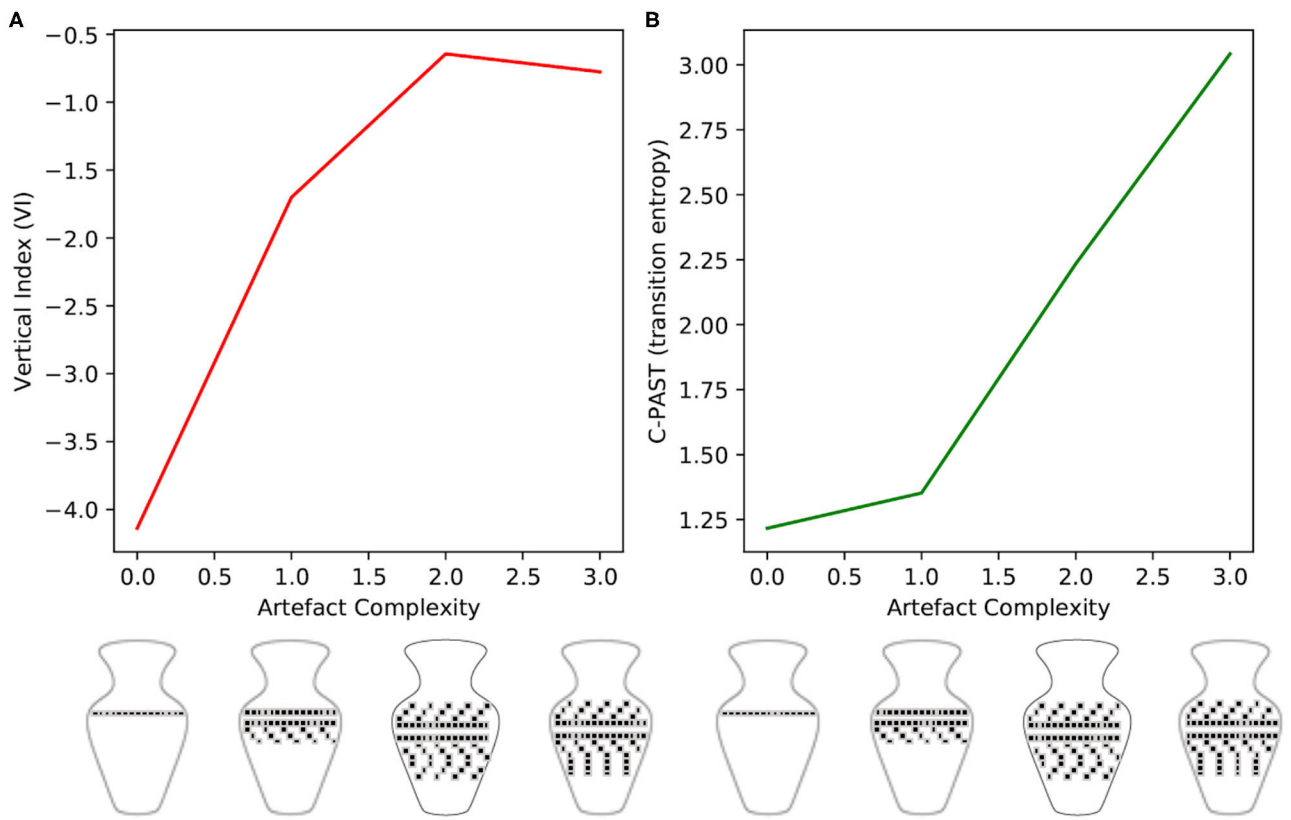



**FIGURE 3 | (A)** The effect of artefact complexity on the vertical index (Vi) measure of scan paths. Consistent with the empirical findings of Criado-Boado et al. (2019), we find that artefactual complexity positively correlates with Vi **(B)**. The effect of artefact complexity on the C-PAST measure of learned generative models. As described in the main text, this measure quantifies the entropy of the motif-transition parameters "B2," which are learned over the course of the 100 trials. These results demonstrate that artefact complexity correlates positively with C-PAST, highlighting the symmetry between environmental complexity and model complexity.

one from allowing learning in other parameters so as to get a richer measure of C-PAST (e.g., entropy of A and B parameters).

## Results Experiment 2

The goal of experiment 2 was to explore the impact of different C-PASTs on cognitive task performance in a novel cognitive task. Note that the learning only happens in experiment 1. This means that we simply import the trained or learned parameters into the experiment two without letting the model further learn within the context of experiment 2. Accordingly, experiment 2 is not a typical transfer learning experiment. However, the proposed setup is ready for bone fide transfer learning simulations as future work could allow for learning, and thus study the effect of transferred learning on learning and task performance. Here, we only focused on the effect of prior learning on novel task performance. Experiment 1, which could be viewed as a "training," or learning experiment was a visual foraging task. Experiment 2 is a simple visual classification task where the agent is presented with a predetermined series of cut-outs of certain shapes and must select the shape that matches the cut-out (see **Figure 5**). We simulated the task under the four different C-PASTs acquired in experiment 1. These were acquired through the exposure to the four different levels of decoration complexity on the vase. We presented the same predetermined series of cut-outs to all agents. We then recorded hits and non-hits over 100 trials, or series of 100 cut-outs. The agent received no feedback on its answer, meaning that no further learning took place in experiment 2.

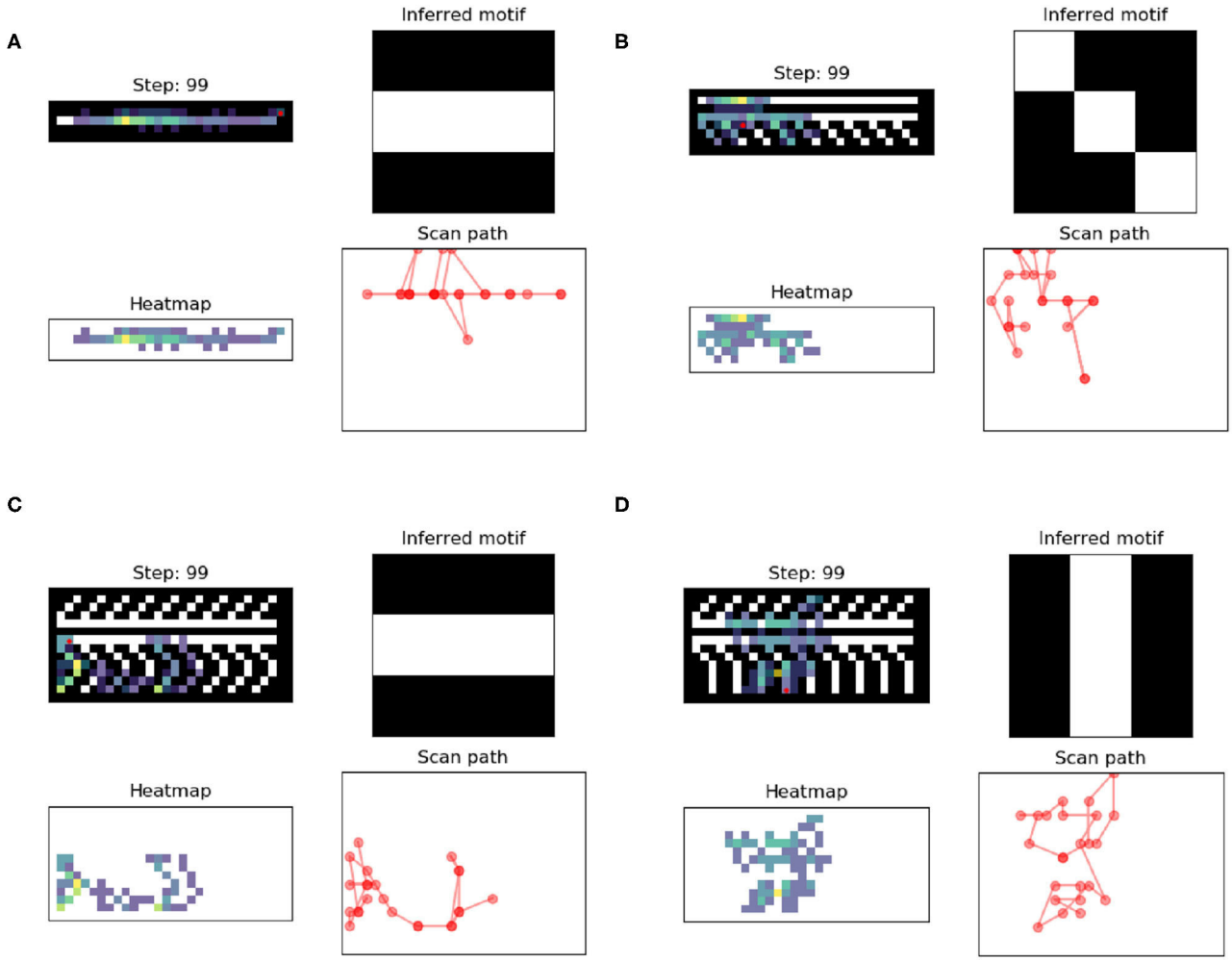



**FIGURE 4 |** For each panel [**(A–D)**, or complexity 0 to 3], the upper left quadrant represents the heatmap and the final location (red dot), with the motifs superimposed at the end of the 99 trials; the lower left quadrant represents the heatmap alone at the end of the 99 trials; and the lower right quadrant represents the scan path over 99 trials; the upper right quadrant represents the final motif perceived. The heatmap represents the paths and the number of times the agent moved over a location. The more often a location was gazed upon, the yellower it became (gradient from blue to yellow). MP4 versions of the experiment can be found at https://github.com/BerenMillidge/MaterialCulture.

### The Model for Experiment 2

We use the same model as in experiment 1, but with a single level of parametrisation. Hidden states correspond to the shapes that made up the motifs in experiment 1, and the sensory outcomes are the cut-outs. Accordingly, the transition matrices (B parameters) are the transition probability mappings between shapes, and the likelihood parameter (A parameter) is a likelihood mapping between cut-outs and shapes. The likelihood mapping is deterministic, meaning that our agent can perfectly sense the shapes and their associated cut-outs (i.e., the agent has a non-noisy sensory access to cut-outs). The mappings for the transition parameters are those that have been learned in experiment 1 (i.e., as level 2 transition probability mappings, or level 2 B parameters B2), and so may contain uncertainty. Crucially, and distinct in this work is that we use a novel method of "likelihood remapping" to ensure that the agent at any point only perceives its local environment—i.e., the central foveated region of the visual.

### Performance

The stimuli we employed in experiment 2 was a series of cut-outs. The task was to select the matching cut-out. We recorded the number of hits and non-hits over 100 trials. **Figure 6** presents the results for the agents having been trained under the four different levels of decoration complexity in experiment 1. Our results show that C-PAST trained under higher levels of cultural complexity leads to increased performance.

## DISCUSSION

This paper presents two computational experiments using active inference. The first was a training simulation wherein an

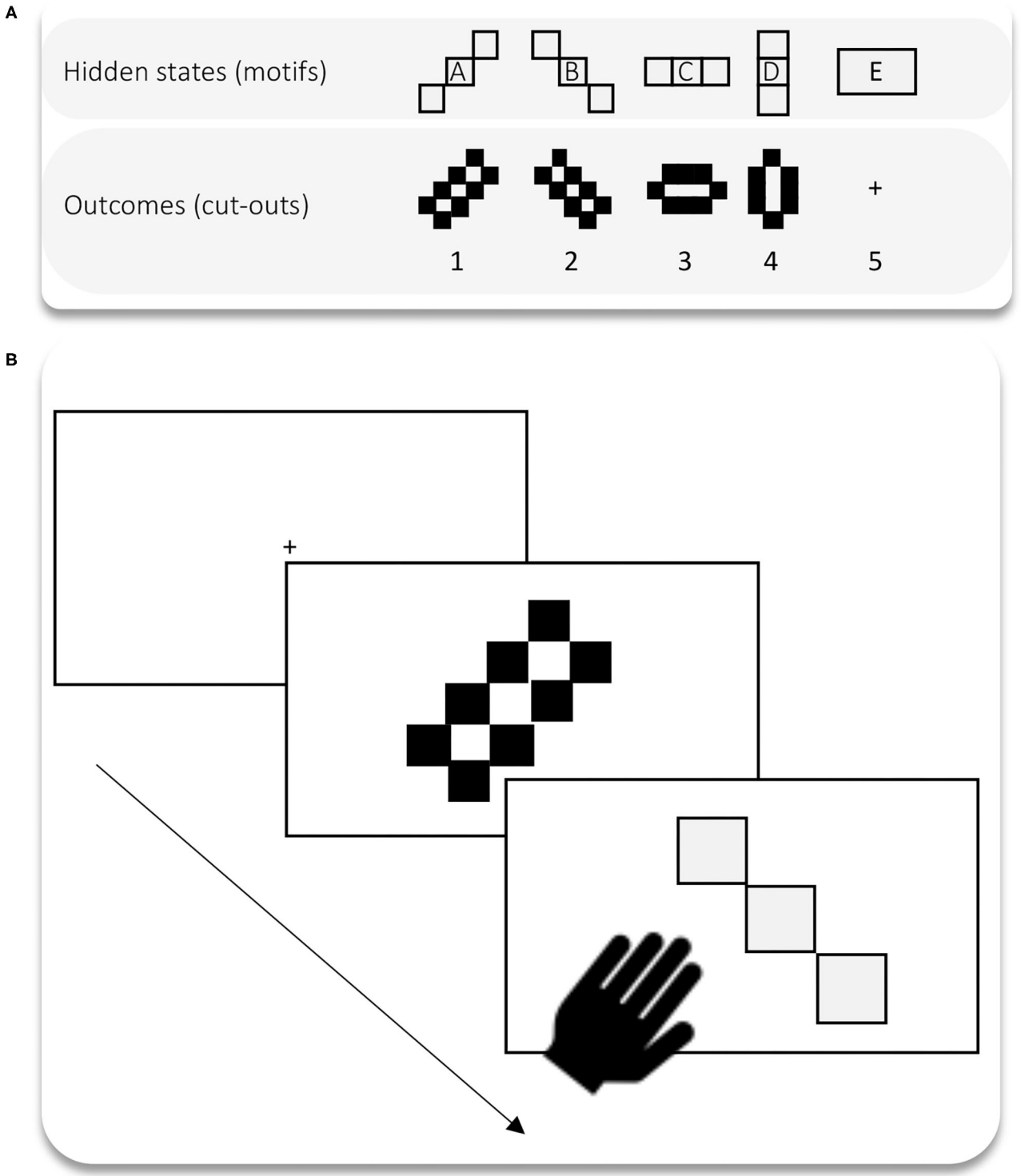



**FIGURE 5 | (A)** States and outcomes are the shapes and the cut-outs respectively. There are no preferences for specific cut-outs; only aversion for the outcome that corresponds to the "+". This is to make sure that the simulated agent always makes a decision across the 100 trials. **(B)** The task wherein the agent is presented with a series of cut-outs (outcomes) and has to infer what shape should fit in. We ran 100 trials where the agent is presented with a blank display followed by a cut-out. A single trial has three moments: (i) the agent is presented with a target (first slide); (ii) the agent receives the sensory entry, or cue that corresponds to the cut-out (second slide); (iii) the agent infers and selects the motif that matches the cut-out (slide 3).

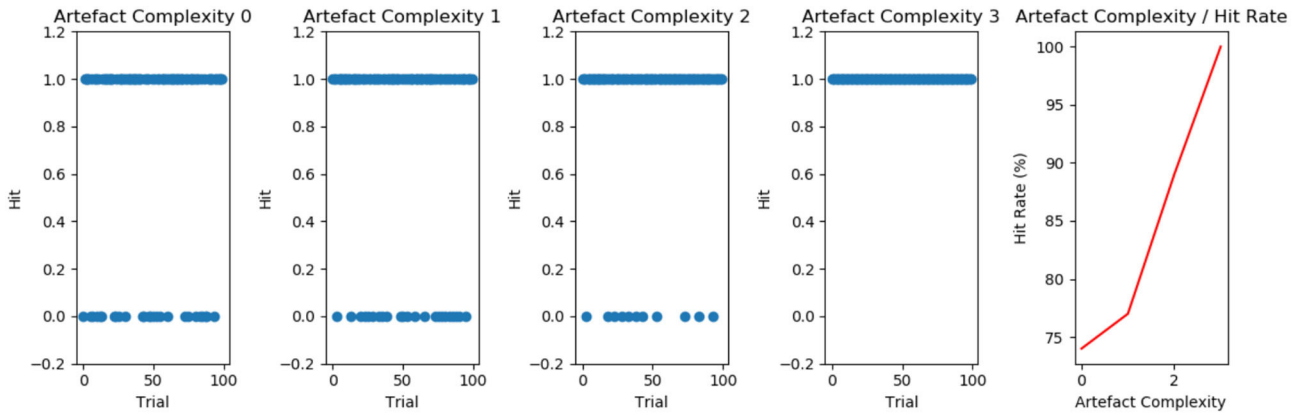



**FIGURE 6 |** From left to right: Hits (1.0) vs. non-hits (0.0) over 100 trials for agent's exposed to different levels of artefact complexity. We find that the accuracy of agent's guesses (percentage of hit-rate) correlates with the level of artefact complexity.

agent could freely explore the decorations on four different vases affording four different levels of decoration complexity. Decoration complexity was defined in terms of the amount of verticality in motifs as well as the number of motifs. We used the decoration complexity to train different culturally patterned attention styles (C-PASTs), which we phenotyped in terms of the entropy of the distribution of their associated parameters. We observed that the different C-PASTs correlated with different virtual indices (Vi). The overall observation was that the increase in complexity correlated with increase in Vi, as observed empirically by Criado-Boado et al. (2019). Complexity and Vi also correlated positively with C-PASTs' entropy. The goal of our simulation study was simply to reproduce the results of Criado-Boado and colleagues using active inference. Because active inference is known as a good computational candidate to account for human behaviour, the main contribution of our simulation is to have provided the beginning of a behaviourally plausible explanation for the computation that may underwrite the results of Criado-Boado and colleagues. The explanation for the correlation we simulated is simple: the more complex the stimuli, the more exploration there is, and the more exploration there is, the more transitions are observed and therefore the more dispersion there is in the B2 parameters (i.e., the mappings are less deterministic). Cultural complexity thus has the consequence of "loosening" the learning of transitions among cultural motifs, and so renders learning more flexible (i.e., opens the agent to exploring novel shapes), which is a phenomenon discussed in relation to creativity (Van de Cruys and Wagemans, 2011; Veissière et al., 2020). There is more complexity and variety in the experienced transitions amongst hidden states, or more elaborate hierarchical structure, which in turn facilitates learning more complex and varied models of the world. In experiment 2, we simulated a simple visual classification task in which we reused the C-PAST trained in experiment 1. Here again, increased flexibility in learning prove useful. We measured success rate (hits non-hits) in a simple visual discrimination task under each C-PAST. The overall observation was that C-PASTs acquired during the exploration of more complex artefacts lead to better performance in the discrimination task.

Crucially, our results on the relation between cultural complexity and the "loosening" of the learning of transitions among cultural motifs are consistent with archaeological observations. For instance, in neolithic contexts, it has been observed that relatively uniform ceramic decorations increase the diversity of the decoration over time. For example, we can structurally identify Neolithic societies in Central Europe for which oldest phase uniform decorations are in use over a large area. In the following phase, this uniformity dissolves, which correlates with increase in decorations variability. In the case of Linear Pottery, this is associated with spin-offs of individual farmsteads from the central settlement around 5100 B.C. and increased generational independence (Shennan and Wilkinson, 2001). A similar phenomenon can be seen for the large-scale Globular Amphora phenomenon with a broadening of ritual activities around 3,000 BC (Müller, 1996). From 2500 B.C. onwards, cyclical increases and decreases in motif variation are observed for the Bell Beakers, which can be linked to an intentional renewed restriction of cultural diversity usually occurring every 150 years or so. Since comparable changes in diversity are also probable in the Bronze Age (cp. Staniuk, 2020), we should be able to identify a fundamental phenomenon for illiterate societies. The learning changes observed in the simulations offer at least one of several explanatory patterns for the archaeological observations described in the example.

The purpose of experiment 1 and 2 was to demonstrate the feasibility of an active inference based archaeological study of the effect of material culture on cognition. In future work, we plan on using the computational paradigm developed here to test empirically the correlations observed in our simulated experiment. Even though we used very simple tasks in experiments 1 and 2 for illustrative purposes, nothing prevents us from designing more complex simulation scenarios that can be used to model participant's performance in richer environments. Indeed, the likelihood remapping strategy we employed in this

paper, because it builds environments based on a single type of hidden states, makes it possible to design complex 2D or 3D training environments and to transfer the learning of the model across different such environments.

Projected iterations of this new experimental paradigm could address at least four important and interlocking issues. The first, and most obvious, is to explore the effects of different material structures and practises on learning and attention. This could be done with contemporary artefacts characteristic of different culture, which we could use to test cross-cultural variations in visual attention styles. Such future studies should be informed by studies on cultural differences in physical objects perception (e.g., Masuda and Nisbett, 2001, 2006; Kitayama et al., 2003; Ishii et al., 2014). The second is to explore learning and transmission in whole populations of active inference agents. The third is to look at how learning that is achieved in one such generation and passed on to another influences the design of the environment itself—the so-called "trans-generational bottleneck" whose importance in the domain of language change has been the subject of much recent experimentation (for a review, see Smith and Kirby, 2008). Here, there is an opportunity to confront the real historical record with predictions made on the basis of the simulations. The fourth—and potentially the most revealing—would be to explore the principal dimensions along which variations in material culture and patterned practises impact learning and attention, using this to drive new (more functionally revealing) ways of grouping and taxonomising the real socio-historical record. For example, we predict that important variations will flow from the way different material designs manipulate sensory surprise at different levels of abstraction and processing.

Summing up, we have described a new experimental pipeline for exploring links between active inference and changing cultural complexity. These links are, we hypothesise, mediated by changing patterns of attention—patterns that can be trained and enforced by the structural and decorative complexity of the objects we encounter. In future work using this pipeline, we hope to discover more of the hidden variables and deep guiding principles linking material culture to changing patterns of thought and reason.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

All authors contributed to the redaction of the manuscript. AT, BM and AC contributed to the development and writing of the code.

## FUNDING

Researchers on this article were supported by an Australian Laureate Fellowship project A Philosophy of Medicine for the 21st Century (Ref: FL170100160) and by a Social Sciences and Humanities Research Council doctoral fellowship (Ref: 752-2019-0065) (AC), by a PhD studentship from the Sackler Foundation and the School of Engineering and Informatics at the University of Sussex (AT); by an EPSRC PhD Studentship (BM), by a GAIN-Xunta de Galiza Groups of Excellence 2020 (FC-B), and by Horizon 2020 European Union ERC Advanced Grant XSPECT - DLV-692739 (AC). AT is grateful to the Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.729665/full#supplementary-material

## REFERENCES

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again.* Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1552.001.0001

Constant, A., Clark, A., Kirchhoff, M., and Friston, K. J. (2020). Extended active inference: constructing predictive cognition beyond skulls. *Mind Lang.* 1–22. doi: 10.1111/mila.12330

Constant, A., Ramstead, M. J., Veissière, S. P., and Friston, K. (2019). Regimes of expectations: an active inference model of social conformity and human decision making. *Front. Psychol.* 10:679. doi: 10.3389/fpsyg.2019.00679

Criado-Boado, F. (2014): "Archaeologies of space: an inquiry into modes of existence of Xscapes," in *Paradigm Found: Archaeological Theory – Present, Past and Future. Essays in Honour of EvŽen* Neustupný, eds K. Krisitansen, L. Smejda, and J. Turek (Oxford: Oxbow Books), 61–83. doi: 10.2307/j.ctvh1dpc1.9

Criado-Boado, F., Alonso-Pablos, D., Blanco, M. J., Porto, Y., Rodríguez-Paz, A., Cabrejas, E., et al. (2019). Coevolution of visual behaviour, the material world and social complexity, depicted by the eye-tracking of archaeological objects in humans. *Sci. Rep.* 9:3985. doi: 10.1038/s41598-019-39661-w

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *J. Math. Psychol.* 99:102447. doi: 10.1016/j.jmp.2020.102447

Dennett, D. (1991). *Consciousness Explained.* Boston: Little Brown.

Dennett, D. (1996). *Kinds of Minds.* New York, NY: Basic Books.

FitzGerald, T. H. B., Schwartenbeck, P., Moutoussis, M., Dolan, R. J., and Friston, K. (2015). Active inference, evidence accumulation, and the urn task. *Neural Comput.* 27, 306–328. doi: 10.1162/NECO_a_00699

Fountas, Z., Sajid, N., Mediano, P. A. M., and Friston, K. (2020). Deep active inference agents using Monte-Carlo methods. arxiv [q-bio.NC].arxiv:2006.04176.

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7:598. doi: 10.3389/fnhum.2013.00598

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29 1–49. doi: 10.1162/NECO_a_00912

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P. O., Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev*. 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci*. 6, 187–214. doi: 10.1080/17588928.2015.1020053

Henrich, J. (2015). *The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species, and Making us Smarter*. Oxford: Princeton University Press. doi: 10.2307/j.ctvc77f0d

Heyes, C. M., and Frith, C. D. (2014). The cultural evolution of mind reading. *Science* 344:1243091. doi: 10.1126/science.1243091

Ishii, K., Miyamoto, Y., Rule, N. O., and Toriyama, R. (2014). Physical objects as vehicles of cultural transmission: maintaining harmony and uniqueness through colored geometric patterns. *Pers. Soc. Psychol. Bull*. 40, 175–188. doi: 10.1177/0146167213508151

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern*. 112, 323–343. doi: 10.1007/s00422-018-0753-2

Kaufmann, L., and Clément, F. (2007). How culture comes to mind: from social affordances to cultural analogies. *Intellectica* 46, 221–250. doi: 10.3406/intel.2007.1286

Kitayama, S., Duffy, S., Kawamura, T., and Larsen, J. T. (2003). Perceiving an object and its context in different cultures: a cultural look at new look. *Psychol. Sci*. 14 201–206. doi: 10.1111/1467-9280.02432

Knappett, C. (2005). *Thinking Through Material Culture: An Interdisciplinary Perspective*. Philadelphia, PA: University of Pennsylvania Press. doi: 10.9783/9780812202496

Masuda, T., and Nisbett, R. E. (2001). Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. *J. Pers. Soc. Psychol*. 81:922. doi: 10.1037/0022-3514.81.5.922

Masuda, T., and Nisbett, R. E. (2006). Culture and change blindness. *Cogn. Sci*. 30, 381–399. doi: 10.1207/s15516709cog0000_63

Millán-Pascual, R., Martínez, L. M., Alonso-Pablos, D., Blanco, M. J., and Criado-Boado, F. (2021). Materialidades, espacio, pensamiento: arqueología de la cognición visual. *Trabajos De Prehistoria* 78, 7–25. doi: 10.3989/tp.2021.12262

Millidge, B. (2020). Deep active inference as variational policy gradients. *J. Math. Psychol*. 96:102348. doi: 10.1016/j.jmp.2020.102348

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci*. 10:56. doi: 10.3389/fncom.2016.00056

Müller, J. (1996). The use of correspondence analysis for different kinds of data categories: domestic and ritual Globular Amphorae sites in Central Germany. *Analecta Praehistorica Leidensia* 28, 217–222.

Müller, J., Hinz, M., and, Ullrich, M. (2015): "Bell beakers: chronology, innovation and memory: a multivariate approach," in *The Bell Beaker transition in Europe. Mobility and Local Evolution During the 3rd Millennium BC*, eds M. P. Prieto-Martínez and L. Salanova (Oxford: Oxbow Books), 57–68.

Oswalt, W. H. (1973). *Habitat and technology: The evolution of hunting*. New York, NY: Holt, Rinehart, & Winston.

Parr, T., and Friston, K. J. (2017). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14:20170376. doi: 10.1098/rsif.2017.0376

Pio-Lopez, L., Nizard, A., Friston, K., and Pezzulo, G. (2016). Active inference and robot control: a case study. *J. R. Soc. Interface* 13:20160616. doi: 10.1098/rsif.2016.0616

Prieto-Martínez, P., Cobas-Fernández, I., and Criado-Boado, F. (2003). "Patterns of spatial regularity in late prehistoric material culture styles of the NW Iberian Peninsula," in *Prehistoric Pottery: People, Pattern and Purpose 1156*, ed A. Gibson (Oxford: British Archaeological Reports), 147–188.

Ramstead, M. J., Veissière, S. P., and Kirmayer, L. J. (2016). Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. *Front. Psychol*. 7:1090. doi: 10.3389/fpsyg.2016.01090

Renfrew, C., and Malafouris, L. (eds.). (2010). *The Cognitive Life of Things: Recasting the Boundaries of the Mind*. Cambridge: McDonald Institute for Archaelogical Research.

Shennan, S., and Wilkinson, J. R. (2001). Ceramic style change and neutral evolution: a case study from Neolithic Europe. *Am. Antiquity* 66, 577–593. doi: 10.2307/2694174

Smith, K., and Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. 363, 3591–3603. doi: 10.1098/rstb.2008.0145

Staniuk, R. (2020). *Kakucs-Turjan. Tradition and Practice. Study on Pottery, Chronology and Social Dynamic of the Hungarian Bronze Age*. Bonn: Habelt.

Sterelny, K. (2020). Demography and cultural complexity. *Synthese*. doi: 10.1007/s11229-020-02587-2

Sutton, J. (2002). "Porous memory and the cognitive life of things," in *Prefiguring Cyberculture: An Intellectual History*, eds D. Tofts, A. Jonson, and A. Cavallaro (Cambridge, MA: MIT Press), 130–141.

Tschantz, A., Millidge, B., Seth, A. K., and Buckley, C. L. (2020). Reinforcement learning through active inference. arxiv [cs.LG].arXiv:2002.12636.

Van de Cruys, S., and Wagemans, J. (2011). Putting reward in art: a tentative prediction error account of visual art. *I-Perception* 2, 1035–1062. doi: 10.1068/i0466aap

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., and Kirmayer, L. J. (2020). Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 43:E90. doi: 10.1017/S0140525X19001213

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2021 Constant, Tschantz, Millidge, Criado-Boado, Martinez, Müeller and Clark. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Neurorobotics

PERSPECTIVE
published: 31 January 2022
doi: 10.3389/fnbot.2021.772012

# Enacting Plant-Inspired Robotics

*Jonny Lee* and Paco Calvo**

*Minimal Intelligence Laboratory (MINTLab), University of Murcia, Murcia, Spain*

Plants offer a source of bioinspiration for soft robotics. Nevertheless, a gap remains in designing robots based on the fundamental principles of plant intelligence, rooted in a non-centralized, modular architecture and a highly plastic phenotype. We contend that a holistic approach to plant bioinspiration—one that draws more fully on the features of plant intelligence and behavior—evidences the value of an enactivist perspective. This is because enactivism emphasizes not only features of embodiment such as material composition and morphology, but also autonomy as an important aspect of plant intelligence and behavior. The enactivist sense of autonomy concerns the dynamics of self-producing systems (such as plants) that create a distinction between themselves and a domain of interactions that bear on the conditions of viability of the system. This contrasts with the widespread, but diluted notion of autonomy that merely indicates the independent operability of a system for an arbitrary period. Different notions of autonomy are relevant for soft roboticists, for instance, when evaluating limitations on existing growing robots ("growbots") that take bioinspiration from plants, but depend on a fixed source of energy and material provided by an external agent. More generally, plant-inspired robots serve as a case study for an enactivist approach to intelligence, while, correspondingly, enactivism calls attention to the possibility of non-zoological forms of intelligence embodied in a self-organizing, autonomous system.

**Keywords: soft robotics, embodied robotics, plant intelligence and behavior, enactivism, autonomy, growbots**

OPEN ACCESS

***Edited by:***
*Andy Clark,*
*University of Sussex, United Kingdom*

***Reviewed by:***
*Isabella Fiorello,*
*Istituto Italiano di Tecnologia, Italy*
*Joe Gough,*
*University of Sussex, United Kingdom*

****Correspondence:***
*Jonny Lee*
*jonathan.lee@um.es*
*Paco Calvo*
*fjcalvo@um.es*

***Received:*** *07 September 2021*
***Accepted:*** *30 December 2021*
***Published:*** *31 January 2022*

***Citation:***
*Lee J and Calvo P (2022) Enacting Plant-Inspired Robotics. Front. Neurorobot. 15:772012. doi: 10.3389/fnbot.2021.772012*

## INTRODUCTION

Plants offer a rich source of bioinspiration for soft robotics. Despite progress in selected areas (see Mazzolai et al., 2020, for a mini-review), a gap remains in designing systems based on the fundamental principles of plant intelligence. More "holsitically" plant-inspired robots would inhabit bodies that exhibit a fuller range of plant features, rooted in a decentralized and modular architecture coupled with a highly plastic phenotype (Calvo et al., 2020; Calvo and Trewavas, 2021). In addition to plant-like bodies, realizing key characteristics of plant intelligence, such as flexible and adaptive growth, may require attention to the role of biological autonomy. Given its consideration of embodied features such as material composition and morphology as well as adaptive autonomy, this article indicates that the project of designing more fully plant-like systems forms a fruitful two-way exchange with enactivism (Varela et al., 1991/2017; Noë, 2004; Stewart et al., 2010; Thompson, 2010; Hutto and Myin, 2012; Di Paolo et al., 2017).

The prospect of more holistically plant-inspired robots connects with a general embodied perspective that recognizes the value of intelligent problem-solving via adaptive morphology [as demonstrated, for example, exemplar "passive dynamic walker" by McGeer (1990); for discussion, see Clark, 1997]. Smart embodiment is evidently key to plant intelligence and behavior; for instance, the material and structural properties of plant bodies are adapted to exploit physical constraints

(friction, gravity, and inclination) for growth (as opposed to locomotion) (Lopez et al., 2014; Vandenbrink and Kiss, 2019). Correspondingly, plant-inspired robots indicate alternative means of adaptive embodiment in the form of growing robots or "growbots" (Laschi et al., 2016; Sadeghi et al., 2017; Del Dottore et al., 2019), i.e., systems that *move* by lengthening or extending the surface area of their bodies.

Beyond these basic considerations of embodiment, an enactive perspective also draws attention to a strong sense of autonomy, grounded in the concept of autopoiesis (Vernon, 2010). As such, enactivism can play a heuristic role in drawing attention to strong biological autonomy and reminding us that materials and morphology do not exhaust the possibilities of bioinspiration. As it pertains to plant-inspired robotics, this perspective can be used (among other things) to evaluate limitations on existing growbots, which take bioinspiration from plants, but depend on a fixed source of energy and material provided by an external agent. More broadly, considering autonomy as part of a soft and embodied perspective may serve in the development of holistically plant-like robots, while testing principles of non-animal intelligence and behavior gleaned from applying tools from plant cognitive science/neurobiology (Baluška et al., 2006a,b).

## EXISTING PLANT-INSPIRED ROBOTS AND THE NATURE OF PLANT INTELLIGENCE

Existing bioinspired robots demonstrate the practical value of considering plant capacities for intelligent behavior. Recent advances in material composition, kinematic principles, and morphological features build on plant research. For example, effective adhesive mechanisms have been drawn from examinations of climbing plants, soft spiral grippers from twinning plants (Yang et al., 2020), and grasping-by-coiling behaviors from plant circumnutation—a term coined by Darwin (1875) that refers to the helical movements created by growing tips and other plant organs. Moreover, robotic growth via root-like filament deposition has taken inspiration from the plant kingdom (Blumenschein et al., 2020; Fiorello et al., 2020; Mazzolai et al., 2020).

Much of this existing plant-inspired research falls within the field of soft robotics, which is vital for understanding the holistic plant-inspired robotics targeted in this article. By "holistic plant-inspired robotics," we refer to the development of systems that are more fully plant like in their intelligence and behavior (in a sense to be specified shortly), as opposed to merely borrowing a small number of specific materials or gadgets. Soft robotics refers to the design and construction of systems with flexible bodies using compliant materials, often drawing on the properties of living organisms (Kim et al., 2013; Calisti et al., 2017; Thieffry et al., 2017; Rich et al., 2018; Drotman et al., 2021). A common advantage of soft (over hard) robots is greater bodily flexibility and adaptability to the environmental constraints. Soft robotics, in turn, overlaps with "embodied" perspectives, introduced earlier. While soft robotics focuses specifically on the problem-solving potential afforded by compliant materials of the sorts exploited by nature (Trivedi et al., 2008), embodied perspectives more broadly draw insight from the capacities of the adaptive morphology of an organism (Hoffmann and Pfeifer, 2018). In keeping with a soft and embodied perspective, research in plant intelligence indicates the distributed nature of control and processing, where adaptive responsibility is shared between internal signaling channels, the material properties of (soft) organs, and the dynamics of body-environment interactions.

By examining existing plant-inspired robots, we can distinguish between systems that selectively borrow elements of plant design vs. systems based on the fundamental organizing principles of plant intelligence (Frazier et al., 2020). There is a spectrum. However, plant-inspired robotics has hitherto concentrated on a small number of tools for solving certain problems (although see Blumenschein et al., 2020, for instance, on designing more plant-like systems of control). As such, there remain unexplored avenues for engineering systems that manifest the full suite of fundamental features of plant intelligence. Such systems not only contain a few plant-like gadgets, but resemble plants in their basic organization.

Of course, plants exhibit as much variety in their anatomical and physiological details as animals. We should, therefore, remain sensitive to potential diversity in plant intelligence and behavior. Nevertheless, we can identify some generic principles that typify the plant kingdom, much as we can with animals (aardvarks, albatrosses, and alligators share similar centralized neural hardware and locomotion-based sensorimotor competencies, despite their myriad differences). Indeed, attending to the common character underlying plant particularities might help us to appreciate the gaps left by plant-inspired robotics that focuses only on specific bodily gadgets. The key features of plant behavior and intelligence that we take to be instructive for soft roboticists include the following:

**Distributed coordination:** Higher plants are characterized by a highly globalized yet decentralized, i.e., distributed architecture, with replicating modules that consist of branch roots (below ground) alongside leaves and subtended buds (above ground), flexibly distributed to optimize the procurement of energy and mineral resources (Calvo and Trewavas, 2021). The important point, for our purposes, is that plants display highly localized activity, while using feedback and feedforward mechanisms (Calvo and Friston, 2017) to provide stability and flexible responses to achieve organism-level adaptive behavior.

**Movement via growth:** Plants move by growth rather than locomotion (Darwin and Darwin, 1880). In animals, growth principally concerns the development of the organism as it matures and is relatively determined. In plants, growth is associated with the continuous, dynamic interaction of the organism with the environment, throughout its life, and is highly plastic. It is primarily characterized by the extension from the tip of the body (apical extension) and length change, allowing organisms to move through spatially constrained environments and adopt three-dimensional structures. Growth, thus, closely overlaps with "remodeling" of a plant, changing its material properties, and "morphogenesis," changing its shape, to adaptively act within its dynamic environment (Del Dottore et al., 2018). Notably, as an efficient strategy for movement, growth

is found across scales of natures and different kingdoms—for example, in fungal hyphae as well as networks of neurons—and is associated with the flexible exploration of three-dimensional (3D) space in a non-deterministic body (Blumenschein et al., 2020).

**Neural-like properties:** Plants lack neurons. Nevertheless, growing research highlights related molecular-level functional similarities between animal and plant substrates (Baluška and Levin, 2016; Miguel-Tomé and Llinás, 2021). One example is the fact that plants possess neurotransmitters [acetylcholine, glutamate, dopamine, histamine, noradrenaline, serotonin, and gamma-aminobutyric acid (GABA)], some of which appear to play roles analogous to those in animals (Baluška and Mancuso, 2009a; Baluška, 2010). Another example is the capacity for plant cells to produce electric potentials and exploit auxin-secreting neuron-like plant synapses (Baluška and Mancuso, 2009b). Electrical signals are transmitted along vascular conduits via networks of phloem, xylem, and cambium, again highlighting the importance of the vascular system for whole-body integration (Baluška et al., 2006).

**Swarm intelligence:** Swarm intelligence refers to the activity of the decentralized group of individuals that collectively results in the emergence of adaptive behavior. Examples include bird flocking, microbial organization, ant colony coordination, and fish schooling. Research suggests that swarm intelligence might apply to the plant roots too: local interactions between relatively simple components (root tips) result in the emergent functionality. For instance, Ciszak et al. (2012) argue that coordinated activity among individual root apices, which change in growth direction produces their episodic patterns of coordinated activity, resulting (collectively) in resource optimization.

Through their modular architecture within a highly plastic phenotype, plants engage in a range of flexible and information-sensitive capacities. Commonly observed capacities include perception, communication, kin recognition, decision-making, anticipation, learning, risk sensitivity, and mimicry (Calvo, 2016; Segundo-Ortin and Calvo, 2021). Plants, thus, display remarkably intelligent behaviors without the need for a central control organ.

## Enacting Bioinspiration

As our discussion so far suggests, designing systems that are more fully plant-like accords with soft robotics and a broader embodied perspective. One reason for this emphasis on soft bodies and smart morphology is that plant intelligence lacks the sort of organization and architecture modeled by symbolic, language like, or more explicitly deliberative architectures (Newell and Simon, 1976; Pylyshyn, 1984). Research in plant intelligence, for instance, indicates the distributed nature of control, where adaptive responsibility is shared between local responses, internal long-distance signaling mechanisms, the material properties of organs, and the dynamics of body-environment interactions (recalling the "principle of ecological balance," Pfeifer and Scheier, 1999). More fully plant-like robots will exploit similar means for adaptive behavior through principles of the smart embodiment such as sensorimotor coupling with soft bodies, and decentralized control (Linson and Calvo, 2020; Calvo and Trewavas, 2021).

Enactivism stresses the role of an adaptive embodiment for intelligence and behavior and, thus, coheres with other soft and embodied perspectives, but additionally centers the role of "autonomy" and "adaptivity" (Froese and Ziemke, 2009), based on the conviction of strong continuity between life and mind (Varela et al., 1991/2017; Thompson, 2007). As with all organisms, such adaptive autonomy plausibly plays an important role in plant intelligence and behavior, as we shall see. We contend, therefore, that an enactive perspective on plant bioinspiration serves as a heuristic for drawing attention to the contribution of soft materials and morphology to plant intelligence as well as ask us to consider the role of adaptive autonomy. On the flipside, plant bioinspiration offers enactivism a case study for exploring the possibility of engineering more fully agential systems.

Enactivism refers to a family of theories that share historical roots and central tenets, but either diverge in significant ways or otherwise stress different aspects of cognition (Ward et al., 2017). For present purposes, the important aspect of enactivism, as we intend it, is that it emphasizes not only: (1) agent-environment coupling and the importance of bodily morphology for intelligent action, in keeping with other embodied approaches, but also the role of (2) *autonomy* (Varela et al., 1991/2017; Thompson, 2007). Autonomy is here defined as a kind of recursive process of production, in which a system is constituted by a network of processes that recursively depend on each other to generate the processes themselves, and constitute the system as a unity individuated from its environment. To quote Thompson, "an autonomous system is a self-determining system, as distinguished from a system determined from the outside or a heteronomous system" (Thompson, 2007, p. 37). For brevity, we focus on basic metabolic or autopoietic autonomy (Ruiz-Mirazo and Moreno, 2004), i.e., the capacity of a system to reproduce and maintain itself physically. However, enactivists often recognize other forms of autonomy (e.g., neurological, immunological, sensorimotor). Robotics and plant research may benefit from attending to these other forms of autonomy, which find a parallel in the plant kingdom. For instance, in addition to "phytoneural" (Calvo et al., 2017) and sensorimotor behavior, we would do well to examine research in plant immunology (Jones and Dangl, 2006; Li et al., 2020).

Complementing the basic idea of autonomous constitution is the idea that a truly autonomous system is "precarious" — it must actively work to ensure its continued existence. This links autonomy with *adaptivity* (Di Paolo, 2005; see also De Jesus, 2018). Contemporary enactivism places great emphasis on adaptivity—the capacity of the system to actively modify its relationship to the environment in a manner that facilitates its persistence (Di Paolo, 2005; Di Paolo and Thompson, 2014). Marrying autonomy with adaptivity, we get "adaptive autonomy" (Barandiaran, 2002, 2004; Barandiaran and Moreno, 2008; Thompson and Stapleton, 2009), i.e., the notion of a system that regulates its interactions with the world, thereby managing its conditions for viability (the conditions under which it persists as a distinct system). This creates a kind of interdependence

between the interaction of a system and its environment and the persistence of that system; actions of a system and its constitution are intertwined.

Although autonomy for enactivists is, strictly speaking, an all or nothing phenomenon—with living systems as the only known instance of an unequivocally autonomous system—we can still think of robots as more or less engineered in relation to enactivist principles. This is because the design of such systems may more or less emphasize autonomy as an important ideal and guiding heuristic (in addition to the importance of morphology and body-environment coupling, shared with other embodied perspectives). Three considerations are worth bearing in mind here. The first is that even embodied robots that are typically thought of as autonomous because they can operate independently for certain durations do not necessarily meet all the requirements for full autonomy in the enactivist sense (Froese and Ziemke, 2009). The second is that even if one falls short of designing a fully autonomous system, autonomy can still function as a model criterion. Finally, a focus on autonomy will produce different results depending on whether research of an individual is animal- or plant-inspired; autonomous growbots may meet different criteria from "locobots" because of their architectural and morphological idiosyncrasies (for a related discussion on the specificity of "organismoid embodiment," see Vernon, 2010).

Autonomy (as well as adaptivity) is argued to be a crucial determiner of genuine agency. We can unpack agency, from an enactivist perspective, in terms of an autonomous organization that adaptively manages its coupling to the environment and, thus, contributes to sustaining itself (Barandiaran et al., 2009). A more exact definition of "basic autonomy" (which slightly diverges from the traditional formulation in terms of autopoiesis) is provided by Ruiz-Mirazo and Moreno: "the capacity of a system to *manage* the flow of matter and energy through it, so that it can, at the same time, regulate, modify, and control: (i) internal self-constructive processes and (ii) processes of exchange with the environment. Thus, the system must be able to generate and regenerate all the constraints—including part of its boundary conditions—that define it as such, together with its own particular way of interacting with the environment" (Ruiz-Mirazo and Moreno, 2004, p. 240. Original emphasis).

An interesting consequence of the enactivist perspective is that relatively "simple" organisms (including all the higher plants) exhibit genuine agency, whereas robots capable of completing complex information-processing tasks typically do not. Even embodied robots with tight perception-action coupling, though perhaps exhibiting agent-like behavior, do not possess intrinsic agency unless such coupling arises from fulfilling one of that requirements of the system for continued survival (Barandiaran et al., 2009; Stapleton, 2016). In short, enactivism provides relevant perspectives for robotic design concerned with the genuine agency, rooted in the biological processes that are not exclusive to animals. Again, it is important to stress the contrast between the concept of autonomy outlined here and one invoked in many areas of robotics (for discussion on the varied of "autonomy" in robotics, see Smithers, 1997). For example, an "autonomous system" often refers to a robot with the mere capacity to self-manage for some extended period (arbitrarily benchmarked) without human supervision.

Take growth in plant-inspired robotics as a case study (Del Dottore et al., 2018). Enactivism provides the tools to assess the limitations of existing growbots, given its emphasis on homeostatic autonomy (Froese and Ziemke, 2009). Existing robots are capable of growth via root-like appendages, providing novel forms of movement (Sadeghi et al., 2017). Recent examples of effective robotic growth include soft pneumatic robots that achieve directed growth through the pressurization of an inverted thin-walled vessel coupled with controlled asymmetric lengthening, displaying a remarkable ability to move through constrained spaces (Hawkes et al., 2017). However, all the existing forms of plant-inspired roots depend on a prefixed store of energy and matter. Recent pressure-driven robots depend on stored material within a "base station" — a fixed spool of polyethylene tubing provides the material for pressure-driven eversion, i.e., turning inside out—and externally provided source of liquid or air pressure (Hawkes et al., 2017). From an enactivist perspective, a more genuinely autonomous robot actively seeks out and metabolizes all the material for growth in its environment and uses this process to aid its persistence as an individuated system. There are existing robots with artificial digestive systems that seek out energy sources, process them, and egest waste (Melhuish et al., 2006; Ieropoulos et al., 2010). Ecobot-II and -III convert biomass into energy using onboard microbial fuel cells with oxygen cathodes. However, these robots still require an external source to supply key materials.

Moving forward, more truly autonomous growbots—that are plant like in not only their material composition and morphology, but in their adaptive autonomy—will not only self-direct and self-manage in the manner of existing so-called "autonomous" robots (free from direct human management), but will actively seek out the requirements for fulfilling the conditions of their own persistence. This may also be relevant in examining limitations in the *amount* of growth and degree of control possible in existing growbots compared with plants, given their dependence on an external source (Hawkes et al., 2017).

## Value of Plant-Inspired Robots

In addition to any generic benefits afforded by an enactivist perspective—for example, see Smithers (1997) on the role of autonomy for navigating unpredictable environments and Lowe and Kiryazov (2014) on the role of autonomy for cognitive-affective processes—designing robots that are more fully plant like in their material composition, morphology, and autonomous control promises some particular advantages for soft robotics. Obviously, autonomous plant-like robots allow us to test the possibilities of what forms intelligence might assume by taking inspiration from a non-zoological branch on the tree of life. They may also allow us to better test existing theories within plant cognitive science/neurobiology, adopting a "synthetic methodology," i.e., understanding a phenomenon by building physical systems that simulate aspects of the phenomenon (Pfeifer et al., 2008).

Robots exhibiting more plant-like bodies as well as stronger autonomy also promise practical benefits. These benefits would

build upon (but potentially surpass) the advantages of existing plant-inspired robots. This includes the fact that plants display high levels of fault tolerance, with catastrophic damage less likely given the absence of system-critical centralized organs as well as the ability to acquire energy and material in proportion to the demands of growth (a function of their adaptive autonomy). In other words, plants have extensive redundancy built-in to their basic organization. Such a strategy can minimize existential risk (no single root is essential), but it also provides novel ways to reach new locations that have advantages over locomotion (e.g., navigating a hard surface by growing through small cracks). There is also the broad principle that engineering an intelligent system via many "not-so-smart" parts—via principles of swarm intelligence—is often optimal given the cost/risk involved. This is especially relevant, for instance, when designing expensive systems for space exploration (Mehling et al., 2006; Wooten and Walker, 2015; Gallentine et al., 2020).

Designing robots with reference to a more complete suite of plant features including stronger autonomy—thus, has the potential to produce relatively low-cost systems which can be deployed with little configuration and that will actively build themselves while exploring and adapting to their environment with little or no external management. This could have serious implications for space exploration, rescue operations, and medical procedures (see also Blumenschein et al., 2020). Plant-inspired robotics, thus, corroborates the dictum that embodied perspectives both offer theoretical insight into the principles of biological intelligence and are of practical value in the design of adaptive systems (Pfeifer et al., 2008).

To summarize, we suggest there are at least four (overlapping) reasons to consider the design of more holistically plant-inspired robots with strong autonomy as a guiding heuristic:

- To uncover novel forms of robotic design (e.g., "is it possible for a robot to solve problem *x* using a plant-like strategy?").
- To exploit unique advantages of plant organization for overcoming real-world tasks (e.g., "can plant-like growth afford special benefits for exploring non-terrestrial planets?").
- To test theories in plant cognitive science/neurobiology (e.g., "can we build a robot with a mechanism analogous to the one we think underlies plant behavior?").
- To engineer robots that exhibit autonomous, decentralized intelligence as proof of concept for what forms intelligence can take (e.g., "what forms of intelligence are possible to engineer and how similar are these to existing organisms?).

Of course, soft roboticists are already sensitive to some of these considerations, some of the time. As such, recognizing the possibility of more holistically plant-like robots partially serves as a tool to deepen and develop existing trends. Equally, if the preceding discussion is correct, too little attention has been paid to the possibility of genuinely autonomous systems, and the use of strong autonomy as a heuristic to develop more fully plant like (and other autonomous) robots, e.g., robots with more genuinely plant-like growth properties.

Our discussion has explored a two-way relationship between enactivism and the design of more plant-like robots. Enactivism helps us attend to the possibility of looking to plants and other non-zoological sources of inspiration, emphasizing the coupling of adaptive morphology with strong autonomy across the tree of life, while the practical success of plant-inspired robots reinforces a postcognitivist perspective (Heras-Escribano, 2019) on the diverse forms intelligence can take (Linson and Calvo, 2020).

## CONCLUSION

This article has only begun to unpack the relationship between plant bioinspiration and enactivism. It is apparent, however, that plants offer a rich source of insight for future developments that overlaps with an enactivist perspective and should not be ignored in favor of purely zoological inspiration. Attention to principles of strong autonomy (as exhibited by plants), in conjunction with novel forms of plant-like materials and morphology, might prove beneficial to plant-inspired robotics. It can also serve to assess the limitations of existing plant-inspired robots such as growbots. More broadly, we indicated that an enactive perspective on plant bioinspiration contributes to ensuring that soft robotics is a productive field that generates theoretical insights as well as practical benefits with quantitative advantages. Future research should examine the overlap between the design of more autonomous plant-inspired robots and existing attempts to develop genuinely life-like systems (Kriegman et al., 2020) as well as other postcognitivist perspectives toward plant bioinspiration such as ecological psychology (Frazier et al., 2020). Finally, in addition to issues pertaining to growth and growbots discussed in this article, work on plant-inspired robotics should investigate the potential of development as a key element in more fully plant-like systems, given the significant role of development in plant adaptive behavior (Segundo-Ortin and Calvo, 2021).

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

## AUTHOR CONTRIBUTIONS

PC conceived the original idea. PC and JL developed the main concepts. JL wrote the outline and manuscript in consultation with PC. Both authors provided feedback and helped shape the manuscript.

## FUNDING

This research was supported by the Office of Naval Research Global (Award # N62909-19-1-2015) to PC and the Juan de la Cierva Fellowship from Ministerio de Ciencia e Innovación del Gobierno de España (Award # FJC2019-041071-I) to JL.

## ACKNOWLEDGMENTS

The authors would thank Xabier Brandiaran for his helpful comments on an earlier version of this manuscript.

## REFERENCES


Baluška, F. (2010). Recent surprising similarities between plant cells and neurons. *Plant Signal. Behav.* 5, 87–89. doi: 10.4161/psb.5.2.11237

Baluška, F., and Levin, M. (2016). On having no head: cognition throughout biological systems. *Front. Psychol.* 7, 902. doi: 10.3389/fpsyg.2016.00902

Baluška, F., and Mancuso, S. (2009b). "Plants and animals: convergent evolution in action?," in *Plant-Environment Interactions*, ed F. Baluška (Berlin; Heidelberg: Springer), 285–301. doi: 10.1007/978-3-540-89230-4_15

Baluška, F., Mancuso, S., and Volkmann, D. (eds.). (2006a). *Communication in Plants—Neuronal Aspects of Plant Life*. New York, NY: Springer. doi: 10.1007/978-3-540-28516-8

Baluška, F., Volkmann, D., Hlavacka, A., Mancuso, S., and Barlow, P. W. (2006b). "Neurobiological view of plants and their body plan," in *Communication in Plants: Neuronal Aspects of Plant Life*, eds F. Baluška, S. Mancuso, and D. Volkmann (Berlin; Heidelberg: Springer), 19–35. doi: 10.1007/978-3-540-28516-8_2

Baluška, S., and Mancuso, S. (2009a). Deep evolutionary origins of neurobiology. *Commun. Integr. Biol.* 2, 1–2. doi: 10.4161/cib.2.1.7620

Barandiaran, X. E. (2002). *Adaptive behaviour, autonomy and value systems. Normative function in dynamical adaptive systems* (MSc thesis). University of Sussex, Brighton, United Kingdom.

Barandiaran, X. E. (2004). "Behavioral adaptive autonomy. a milestone on the Alife route to AI?," in *Artificial life IX: proceedings of the Ninth International Conference on the Simulation and Synthesis of Artificial Life*, eds J. Pollack, M. A. Bedau, P. Husbands, T. Ikegami, and R. A. Watson (Cambridge, MA: MIT Press), 514–521.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* 17, 367–386. doi: 10.1177/1059712309343819

Barandiaran, X. E., and Moreno, A. (2008). Adaptivity: from metabolism to behavior. *Adapt. Behav.* 16, 325–344. doi: 10.1177/1059712308093868

Blumenschein, L. H., Coad, M. M., Haggerty, D. A., Okamura, A. M., and Hawkes, E. W. (2020). Design, modeling, control, and application of everting vine robots. *Front. Robot. AI* 7, 548266. doi: 10.3389/frobt.2020.548266

Calisti, M., Picardi, G., and Laschi, C. (2017). Fundamentals of soft robot locomotion. *J. R. Soc. Interface* 14, 20170101. doi: 10.1098/rsif.2017.0101

Calvo, P. (2016). The philosophy of plant neurobiology: a manifesto. *Synthese* 193, 1323–1343. doi: 10.1007/s11229-016-1040-1

Calvo, P., Friston, K. (2017). Predicting green: really radical (plant) predictive processing. *J. R. Soc. Interface* 14:20170096. doi: 10.1098/rsif.2017.0096

Calvo, P., Gagliano, M., Souza, G. M., and Trewavas, A. (2020). Plants are intelligent, here's how. *Ann. Bot.* 125, 11–28. doi: 10.1093/aob/mcz155

Calvo, P., Sahi, V. P., and Trewavas, A. (2017). Are plants sentient? *Plant Cell Environ.* 40, 2858–2869. doi: 10.1111/pce.13065

Calvo, P., and Trewavas, A. (2021). Cognition and intelligence of green plants. Information for animal scientists. *Biochem. Biophys. Res. Commun.* 564, 78–85. doi: 10.1016/j.bbrc.2020.07.139

Ciszak, M., Comparini, D., Mazzolai, B., Baluška, F., Arecchi, F. T., Vicsek, T., et al. (2012). Swarming behavior in plant roots. *PLoS ONE* 7, e29759. doi: 10.1371/annotation/8e6864fc-c4b7-46e7-92b3-80767f4a5d3a

Clark, A. (1997). *Being There*. Cambridge, MA: MIT Press. doi: 10.1371/journal.pone.0029759

Darwin, C. (1875). *The Movements and Habits of Climbing Plants*. London: John Murray. doi: 10.5962/bhl.title.37759

Darwin, C., and Darwin, F. (1880). *The Power of Movement in Plants*. London: John Murray. doi: 10.5962/bhl.title.102319

De Jesus, P. (2018). Thinking through enactive agency: sense-making, bio-semiosis and the ontologies of organismic worlds. *Phenomenol. Cogn. Sci.* 17, 861–887. doi: 10.1007/s11097-018-9562-2

Del Dottore, E., Mondini, A., Sadeghi, A., and Mazzolai, B. (2019). Characterization of the growing from the tip as robot locomotion strategy. *Front. Robot. AI* 6:45. doi: 10.3389/frobt.2019.00045

Del Dottore, E., Sadeghi, A., Mondini, A., Mattoli, V., and Mazzolai, B. (2018). Toward growing robots: a historical evolution from cellular to plant-inspired robotics. *Front. Robot. AI* 5, 16. doi: 10.3389/frobt.2018.00016

Di Paolo, E., Buhrmann, T., and Barandiaran, X. E. (2017). *Sensorimotor Life: An enactive proposal*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198786849.001.0001

Di Paolo, E., and Thompson, E. (2014). "The enactive approach," in *The Routledge Handbook of Embodied Cognition*, ed L. Shapiro (New York, NY: Routledge), 68–78.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenol. Cogn. Sci.* 4, 429–452. doi: 10.1007/s11097-005-9002-y

Drotman, D., Jadhav, S., Sharp, D., Chan, C., and Tolley, M. T. (2021). Electronics-free pneumatic circuits for controlling soft-legged robots. *Sci. Robot.* 6:eaay2627. doi: 10.1126/scirobotics.aay2627

Fiorello, I., Del Dottore, E., Tramacere, F., and Mazzolai, B. (2020). Taking inspiration from climbing plants: methodologies and benchmarks—a review. *Bioinsp. Biomim.* 15, 031001. doi: 10.1088/1748-3190/ab7416

Frazier, P. A., Jamone, L., Althoefer, K., and Calvo, P. (2020). Plant bioinspired ecological robotics. *Front. Robot. AI* 7, 79. doi: 10.3389/frobt.2020.00079

Froese, T., and Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artif. Intell.* 173, 466–500. doi: 10.1016/j.artint.2008.12.001

Hawkes, E. W., Blumenschein, L. H., Greer, J. D., and Okamura, A. M. (2017). A soft robot that navigates its environment through growth. *Sci. Robot.* 2:eaan3028. doi: 10.1126/scirobotics.aan3028

Heras-Escribano, M. (2019). Pragmatism, enactivism, and ecological psychology: towards a unified approach to post-cognitivism. *Synthese* 198, 337–363. doi: 10.1007/s11229-019-02111-1

Hoffmann, M., and Pfeifer, R. (2018). "Robots as powerful allies for the study of embodied cognition from the bottom up," in *The Oxford Handbook 4e Cognition*, eds A. Newen, L. de Bruin, and S. Gallagher (Oxford: Oxford University Press), 841-861. doi: 10.1093/oxfordhb/9780198735410.013.45

Hutto, D. D., and Myin, E. (2012). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262018548.001.0001

Ieropoulos, I., Greenman, J., Melhuish, C., and Horsfield, I. (2010). "EcoBot-III: a robot with guts," in *Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems* (Cambridge, MA: MIT Press), 733–740.

Jones, J. D., and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323–329. doi: 10.1038/nature05286

Kim, S., Laschi, C., and Trimmer, B. (2013). Soft robotics: a bioinspired evolution in robotics. *Trends Biotechnol.* 31, 287–294. doi: 10.1016/j.tibtech.2013.03.002

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Natl. Acad. Sci.* 117, 1853–1859. doi: 10.106/j.tibtech.2013.03.002

Laschi, C., Mazzolai, B., and Cianchetti, M. (2016). Soft robotics: technologies and systems pushing the boundaries of robot abilities. *Sci. Robot.* 1:eaah3690. doi: 10.1126/scirobotics.aah3690

Li, P., Lu, Y. J., Chen, H., and Day, B. (2020). The lifecycle of the plant immune system. *CRC Crit. Rev. Plant Sci.* 39, 72–100. doi: 10.1080/07352689.2020.1757829

Linson, A., and Calvo, P. (2020). Zoocentrism in the weeds? Cultivating plant models for cognitive yield. *Biol. Philos.* 35, 1–27. doi: 10.1007/s10539-020-09766-y

Lopez, D., Tocquard, K., Venisse, J.-S., Legué V., and Roeckel-Drevet, P. (2014). Gravity sensing, a largely misunderstood trigger of plant orientated growth. *Front. Plant Sci. 5*, 610. doi: 10.3389/fpls.2014.00610

Lowe, R., and Kiryazov, K. (2014). "Utilizing emotions in autonomous robots: An enactive approach," in *Emotion Modeling Lecture Notes in Computer Science 2014* (Cham). doi: 10.1007/978-3-319-12973-0_5

Mazzolai, B., Tramacere, F., Fiorello, I., and Margheri, L. (2020). The bio-engineering approach for plant investigations and growing robots. A mini-review. *Front. Robot. AI* 7, 130. doi: 10.3389/frobt.2020.573014

McGeer, T. (1990). Passive dynamic walking. *Int. J. Robot. Res.* 9, 62–82. doi: 10.1007/s10514-006-6574-5

Mehling, J. S., Diftler, M. A., Chu, M., and Valvo, M. (2006). "A minimally invasive tendril robot for in-space inspection," in *Proceedings BioRobotics 2006 Conference* (Pisa), 690–695.

Melhuish, C., Ieropoulos, I., Greenman, J., and Horsfield, I. (2006). Energetically autonomous robots: food for thought. *Auton. Rob.* 21, 187–198. doi: 10.1007/s10514-06-6574-5

Miguel-Tomé, S., and Llinás, R. R. (2021). Broadening the definition of a nervous system to better understand the evolution of plants and animals. *Plant Signal. Behav.* 16, 1927562. doi: 10.1080/15592324.2021.1927562

Newell, A., and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM* 19, 113–126. doi: 10.1145/360018.360022

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Pfeifer, R., Lungarella, M., and Sporns, O. (2008). "The synthetic approach to embodied cognition: a primer," in *Handbook of Cognitive Science*, eds P. Calvo and Gomila A. (Amsterdam: Elsevier), 121–137. doi: 10.1016/B978-0-08-046616-3.00007-4

Pfeifer, R., and Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1984). *Computation and Cognition*. Cambridge, MA: MIT Press.

Rich, S. I., Wood, R. J., and Majidi, C. (2018). Untethered soft robotics. *Nat. Electron.* 1, 102–112. doi: 10.1038/s41928-018-0024-1

Ruiz-Mirazo, K., and Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artif. Life* 10, 235–259. doi: 10.1162/1064546041255584

Sadeghi, A., Mondini, A., and Mazzolai, B. (2017). Toward self-growing soft robots inspired by plant roots and based on additive manufacturing technologies. *Soft Robot.* 4, 211–223. doi: 10.1089/soro.2016.0080

Segundo-Ortin, M., and Calvo, P. (2021). Consciousness and cognition in plants. *WIREs Cogn. Sci.* e1578. doi: 10.1002/wcs.1578

Smithers, T. (1997). Autonomy in robots and other agents. *Brain Cogn.* 34, 88–106. doi: 10.1006/brcg.1997.0908

Stapleton, M. (2016). Enactivism embraces ecological psychology. *Construct. Found.* 11, 325–327. Available online at: http://constructivist.info/11/2/325

Stewart, J. R., Gapenne, O., and Di Paolo, E. A. (eds.). (2010). *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262014601.001.0001

Thieffry, M., Kruszewski, A., Goury, O., Guerra, T. M., and Duriez, C. (2017). "Dynamic control of soft robots," in *IFAC World Congress* (Toulouse).

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Thompson, E. (2010). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Belknap.

Thompson, E., and Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi* 28, 23–30. doi: 10.1007/s11245-008-9043-2

Trivedi, D., Rahn, C. D., Kier, W. M., and Walker, I. D. (2008). Soft robotics: Biological inspiration, state of the art, and future research. *Appl. Bionics Biomech.* 5, 99–117. doi: 10.1155/2008/520417

Vandenbrink, J. P., and Kiss, J. Z. (2019). Plant responses to gravity. *Seminars Cell Dev. Biol.* 92, 122–125. doi: 10.1016/j.semcdb.2019.03.011

Varela, F. J., Thompson, E., and Rosch, E. (1991/2017). *The Embodied Mind: Cognitive science and human experience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/6730.001.0001

Vernon, D. (2010). Enaction as a conceptual framework for developmental cognitive robotics. *Paladyn J. Behav. Robot.* 1, 89–98. doi: 10.2478/s13230-010-0016-y

Ward, D., Silverman, D., and Villalobos, M. (2017). Introduction: the varieties of enactivism. *Topoi* 36, 365–375. doi: 10.1007/s11245-017-9484-6

Wooten, M., and Walker, I. (2015). "A novel vine-like robot for in-orbit inspection," in *Proceedings 45th International Conference on Environmental Systems* (Bellevue, WA), 1–11.

Yang, M., Cooper, L., and Fok, M. P. (2020). Twining plant inspired pneumatic soft robotic spiral gripper with high-birefringence fiber optic sensor. *Optics Express* 28, 35158–35167. doi: 10.1364/OE.408910

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Lee and Calvo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Psychology

HYPOTHESIS AND THEORY
published: 06 April 2022
doi: 10.3389/fpsyg.2022.846931

# Embodied Intelligence: Smooth Coping in the Learning Intelligent Decision Agent Cognitive Architecture

*Christian Kronsted[1,2]*, Sean Kugele[2,3], Zachariah A. Neemeh[1,2], Kevin J. Ryan Jr[4] and Stan Franklin[2]*

[1]Department of Philosophy, University of Memphis, Memphis, TN, United States, [2]Institute for Intelligent Systems, University of Memphis, Memphis, TN, United States, [3]Department of Computer Science, University of Memphis, Memphis, TN, United States, [4]Department of Philosophy, The University of Tennessee, Knoxville, Knoxville, TN, United States

OPEN ACCESS

**Edited by:**
*Inês Hipólito, Humboldt University of Berlin, Germany*

**Reviewed by:**
*Pierre Gander, University of Gothenburg, Sweden*
*Antonio Chella, University of Palermo, Italy*
*Ricardo R. Gudwin, State University of Campinas, Brazil*

***Correspondence:**
*Christian Kronsted*
*christian.k@memphis.edu*

**Specialty section:**
*This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology*

**Received:** *31 December 2021*
**Accepted:** *02 March 2022*
**Published:** *06 April 2022*

**Citation:**
*Kronsted C, Kugele S, Neemeh ZA, Ryan KJ Jr and Franklin S (2022) Embodied Intelligence: Smooth Coping in the Learning Intelligent Decision Agent Cognitive Architecture. Front. Psychol. 13:846931. doi: 10.3389/fpsyg.2022.846931*

Much of our everyday, embodied action comes in the form of smooth coping. Smooth coping is skillful action that has become habituated and ingrained, generally placing less stress on cognitive load than considered and deliberative thought and action. When performed with skill and expertise, walking, driving, skiing, musical performances, and short-order cooking are all examples of the phenomenon. Smooth coping is characterized by its rapidity and relative lack of reflection, both being hallmarks of automatization. Deliberative and reflective actions provide the contrast case. In Dreyfus' classic view, smooth coping is "mindless" absorption into action, being in the flow, and any reflective thought will only interrupt this flow. Building on the pragmatist account of Dewey, others, such as Sutton, Montero, and Gallagher, insist on the intelligent flexibility built into smooth coping, suggesting that it is not equivalent to automatization. We seek to answer two complementary challenges in this article. First, how might we model smooth coping in autonomous agents (natural or artificial) at fine granularity? Second, we use this model of smooth coping to show how we might implement smooth coping in artificial intelligent agents. We develop a conceptual model of smooth coping in LIDA (Learning Intelligent Decision Agent). LIDA is an embodied cognitive architecture implementing the global workspace theory of consciousness, among other psychological theories. LIDA's implementation of consciousness enables us to account for the phenomenology of smooth coping, something that few cognitive architectures would be able to do. Through the fine granular analysis of LIDA, we argue that smooth coping is a sequence of automatized actions intermittently interspersed with consciously mediated action selection, supplemented by dorsal stream processes. In other words, non-conscious, automatized actions (whether learned or innate) often require occasional bursts of conscious cognition to achieve the skillful and flexible adjustments of smooth coping. In addition, never-conscious dorsal stream information and associated sensorimotor processes provide further online adjustments during smooth coping. To achieve smooth coping in LIDA we introduce a new module to the LIDA cognitive architecture the Automatized Action

Selection sub-module. Our complex model of smooth coping borrows notions of "embodied intelligence" from enactivism and augments these by allowing representations and more detailed mechanisms of conscious control. We explore several extended examples of smooth coping, starting from basic activities like walking and scaling up to more complex tasks like driving and short-order cooking.

**Keywords: smooth coping, automatization, action selection, cognitive architecture, embodied cognition, global workspace theory, LIDA**

## INTRODUCTION

In this article, we develop a conceptual model of smooth coping using LIDA (Learning Intelligent Decision Agent), a hybrid, embodied cognitive architecture implementing the Global Workspace Theory (GWT) of consciousness (Baars, 1988), the perception–action cycle (Neisser, 1976; Freeman, 2002; Fuster, 2004; Cutsuridis et al., 2011), grounded cognition (Harnad, 1990; Barsalou, 1999), appraisal theory (Lazarus, 1991; Roseman and Smith, 2001), long-term working memory (Ericsson and Kintsch, 1995), and other cognitive theories. It aims to be a "unified theory of cognition" (Newell, 1994), taking these and other disparate theories, and uniting them under a single, comprehensive architecture. LIDA is a conceptual and computational architecture that has been used as the basis for software and robotic agents. The current paper is the theoretical overview of how to implement smooth coping in LIDA. Following research will implement formalisms, code agents, and test the agents in various environments. We see this work as a first step toward robot implementation of smooth coping that will fit with current trends in robotics, such as learning by imitation (Bullard et al., 2019).

Smooth coping is the process of skillfully and adaptively acting, typically toward the completion of a task. Smooth coping covers a wide range of skillful behaviors, from those that are relatively basic like breathing or suckling, to those that are learned through painstaking training, as in becoming a pilot (Dreyfus and Dreyfus, 1980). Masterfully driving through traffic, skiing a slope, or running an obstacle course are all classic examples of smooth coping. However, the concept can also include cooking, herding sheep, dancing, tidying up, and many other activities in which it is possible to reach a state of optimized performance. The concept originates in phenomenological philosophy, particularly in the embodied phenomenologies of Heidegger, 1928/2010 and Merleau-Ponty, 1945/2012. Both of these thinkers were reacting against an intellectualized vision of human existence in philosophy and psychology that saw us as essentially epistemic agents geared toward knowing the world. As an alternative, they posited a vision of human existence that was, at its root, pragmatically oriented toward action and movement, and (for Merleau-Ponty) that was based in the agent's embodiment.

In smooth coping the agent is not merely doing disjointed multitasking nor just doing automatized actions. Rather, most of the agent's cognitive processes cohere toward fulfilling one distal intention. We outline how a LIDA agent might achieve smooth coping, and provide three case studies: walking, driving, and short-order cooking (see section "Conclusion"). Importantly, smooth coping in LIDA typically requires a "meshed" combination of conscious, consciously mediated, and never-conscious processes interwoven within a continuing series of cognitive cycles implemented using the Global Workspace Theory of consciousness (Franklin and Baars, 2010). Historically, in the LIDA conceptual model, Action Selection has only been able to choose one, and only one, action at a time. In this paper, we make a significant contribution to the LIDA model by introducing a new sub-module to Action Selection: Automatized Action Selection (AAS). This sub-module allows for concurrent selection of actions—AAS is capable choosing automatized actions in parallel. Furthermore, AAS runs in parallel with the original Action Selection algorithm which continues to choose one action at the time.

We begin by fleshing out recent debates on smooth coping and highlight the meshed nature of cognition supporting it (Christensen et al., 2016; Gallagher and Varga, 2020). We then introduce the LIDA model and the aspects of LIDA relevant to this project. For a more complete overview of LIDA, we recommend reading the tutorial and our two most recent papers (Franklin et al., 2016; Kronsted et al., 2021; Neemeh et al., 2021). We illustrate how smooth coping might take place in a LIDA agent by going through three case studies of increasing complexity: walking alone, driving in traffic, and short-order cooking (see section "Conclusion").

## SMOOTH COPING

Although there has been a recent uptick in debates on smooth coping, the topic can be traced at least back to Aristotle and the notion of *phronesis* (typically translated as "practical wisdom"). Smooth coping debates since their earliest inceptions have typically been tied to culture and sociality—to smoothly maneuver the world is often to do so in rich social cultural contexts (Rietveld and Kiverstein, 2014). Thus, debates on smooth coping cut across discussions in social cognition, anthropology, performance studies, and discussions of "expert performance" (Cappuccio, 2019).

The crossover between motoric and cultural discussions when dealing with smooth coping is especially pronounced when looking at the phenomenological tradition. In the twentieth century, Martin Heidegger introduced the term *Zuhandenheit* in his monumental *Being and Time* (1927). Often translated

as "readiness-to-hand," *Zuhandenheit* refers to a mode of comportment that is pre-reflective and pre-theoretical. When I take something, let us say a tool like a hammer, as ready-to-hand, I am using it rather than reflecting on it. This usage is an embodied know-how rather than theoretical contemplation. Heidegger argued that the Western philosophical tradition focused exclusively on *Vorhandenheit* ("presence-at-hand"), that is, the theoretical comportment. For example, Kant's theory of experience is explicitly aimed at supporting the endeavor of science. This focus on theoretical reason rather than embodied action is something we can see reduplicated in the history of artificial intelligence and robotics. In contrast, Merleau-Ponty, 1945/2012 examined embodiment and action as they dynamically interact with space, time, sexuality, other agents, and other domains. According to Merleau-Ponty, smooth coping is the most fundamental mode of our everyday lives. Years later, Hans Jonas (2001) developed a genetic phenomenology of subjectivity, according to which these basal strata of smooth coping enable higher-order cognitive processes to emerge, similar to contemporary claims of scaffolding. Across thinkers in the phenomenological tradition, we see an emphasis on embodiment in which smooth coping is a basic capacity of cognitive agents as they move through the world. In summary, many phenomenologists take the view that smooth coping forms the basic background of embodied human agency, and that more epistemically oriented, logical, or higher-order processes are less common and are founded against this background.

Building off of the phenomenological tradition, Dreyfus and Dreyfus (1980) developed a cognitive theory of smooth coping based on five stages of skill acquisition. According to their theory, expertise in a skill is characterized by automatization and a lack of higher-order thinking. On this model of smooth coping, experts have habituated their skills within a domain to the point that their movements are fully automatized. This, in turn, is supposed to explain why paying attention to oneself, or deploying higher-order cognitive processes, such as "strategizing," can sometimes be detrimental to performance (Fitts and Posner, 1967; Cappuccio et al., 2019).

In the literature on smooth coping and expert performance, others have followed Dreyfus and Dreyfus and similarly argued that smooth coping in skillful action is a matter of complete automaticity (Papineau, 2013, 2015).

However, the Dreyfus model has in recent years been criticized by a variety of theorists, athletes, and artists, and from a variety of perspectives. For example, Barbara Gail Montero (2010, 2016) demonstrates that to be effective in many sports, the athlete must deploy both automatization and higher-order cognitive processes. Additionally, Montero et al. (2019) demonstrate that the empirical research program claiming that self-attention is detrimental to performance is based on flawed experimental design. Self-attention, monitoring, strategizing, and so forth, are often integrated into the flow of performance, rather than interrupting it.

The point here is that higher-order processes, such as planning, strategizing, monitoring, and so forth, are not always detrimental to expert performance, but on the contrary are often necessary for expert performance and successful smooth coping. Given this insight, smooth coping is often a matter of fluently integrating what some have called "online" (immediate sensory stimuli is needed) and "off-line" (detached from immediate sensory stimuli) cognition (Wilson, 2002). Several theories now propose an integrated web of causality between low-level and higher-order processes in expert performance and smooth coping more generally. Such models include "arch" (Høffding and Satne, 2019), meshed architecture (Christensen et al., 2016, 2019), the dual-process model (Neemeh, 2021), radically meshed architecture (Gallagher and Varga, 2020), and a variety of similar approaches (Bermúdez, 2017; Pacherie and Mylopoulos, 2021).

While these models vary with regards to their commitments, the general gist is the same: both low-level and higher-order cognitive processes are utilized and impact each other during expert performance. For example, automatized non-conscious processes, such as the continual adjustment of posture or dribbling of a basketball, can be impacted by higher-order conscious processes, such as thinking about and realizing the opponent's strategy. A mixed martial arts fighter facing an opponent with a longer reach might strategically try to outsmart their opponent by trying to grapple rather than kicking and punching. Such a higher-order strategic decision in turn impacts how fighters adjust their postures and reconfigure their sensorimotor readiness toward certain action types.

In the literature on dance performance, some phenomenologists have similarly pointed out that even in highly choreographed performances in which one movement brings forth the next, expert dancers must adjust their performances to the particularities of the stage, that night's audience, lighting, air density and humidity, costume malfunctions, and other factors (Bresnahan, 2014). In this same vein, and perhaps even more importantly, the expert dancer (and expert performer in general) must always move in and out of conscious monitoring of the body itself, to adjust in accordance with how the body feels that day (Ravn, 2020).

From these brief examples, we can see that embodied expertise, whether in mundane cases like walking or driving or in highly specialized domains, such as sports and performance, involves a fluent intermixing of various cognitive processes and different levels of awareness (conscious, never-conscious, pre-conscious, pre-reflective). While meshed architecture approaches differ on their commitments to concepts, such as "mental representation" or how to conceptualize the causation between different cognitive mechanisms, it is commonly agreed that smooth coping is not just a matter of automatization. Rather, we frequently utilize and change between various cognitive processes. For example, musicians sometimes report being in a state of complete automatization while simultaneously monitoring their own actions and the actions of fellow musicians. In such a state the musician playing is acting through automatization but they are ready to interject with top-down control at any moment (Høffding, 2019).

Similarly important in discussions of smooth coping and expert performance is the notion of dispositional skill or habit. Here thinkers tend to develop accounts of habits that are strongly inspired by John Dewey's (1922) notion of habit as

a context-sensitive, flexible, disposition to act. Whether working within explicitly anti-representationalist enactive cognitive science (Gallagher, 2020; Segundo-Ortin and Heras-Escribano, 2021) or representationalist cognitive science (Schack, 2004; Sutton et al., 2011; Bermúdez, 2017; Pacherie and Mylopoulos, 2021), there is a general agreement that habit is an important concept in expert performance and smooth coping. Habits in such a view are entrenched through practice but are flexibly adapted to a variety of contexts. Unlike motor programs that are contextually rigid (Ghez, 1985; Neilson and Neilson, 2005), habits are always regulated and finely adjusted by the current context—habits are ways of adaptively being in one's environment (Dewey, 1922).

## THE LEARNING INTELLIGENT DECISION AGENT COGNITIVE ARCHITECTURE

Learning intelligent decision agent is a systems-level cognitive architecture intended to provide a complete and integrated account of cognition (Franklin et al., 2016). Thus, rather than modeling one aspect of mind, the LIDA model aims to be a "unified theory of cognition" (Newell, 1994) capable of modeling human, animal, and artificial minds.[1] Cognition, as it is used here, broadly encompasses every mechanism of mind including (but not limited to) perception, attention, motivation, planning, deliberation, metacognition, action selection, and motor control, as well as the embodiment of all of these activities. "Cognition" then is meant to cover the entirety of the agent's mental life including its embodiment and embodied actions. Within the LIDA framework, "minds" are broadly conceived of as control structures for autonomous agents (Franklin, 1995; Franklin and Graesser, 1997). Here "control structures" (see Newell, 1973) are broadly conceived of as those mechanisms that allow an agent to pursue its agenda. To be an autonomous agent is in part to have an agenda, and to have a mind is to have structures that allow one to pursue that agenda (however simple or complex one's agenda might be). Consequently, autonomous agents are always in the business of answering the question "What should I do next?"

Learning intelligent decision agent is composed of many short- and long-term memory modules, as well as special purpose processors called codelets. While modularity is sometimes seen as a "bad word" in contemporary philosophy of mind, the LIDA model is modular in the sense that it is composed of a collection of independent modules that are constantly performing their designated task. However, it is important to note that the LIDA model is *not* committed to the modularity of brains (Franklin et al., 2013). In fact, the LIDA model makes no claims about brains whatsoever. Thus, the LIDA model can be implemented even by brains that are dynamic and full of neural reuse (Kelso, 1995; Anderson, 2014).

Importantly, the LIDA model implements the Global Workspace Theory of consciousness (Baars, 1988, 2019). An agent typically cannot be aware of everything in its environment (external or internal) and therefore needs to "filter out" the most relevant information. LIDA agents therefore have information regarding the world "compete" for its attention in a module known as the Global Workspace. Whatever structure wins (most typically a coalition of structures) is globally broadcast to every module throughout the model—hence the term "the global broadcast." In this way, the Global Workspace functions as a filter that dictates what information becomes available to the rest of the agent's modules.

In LIDA, sensory stimuli are used to construct both a rich model of the external environment and an internal environment within the module known as the Current Situational Model (CSM). In broad strokes, the CSM creates a model of the world, and different parts of the model are then sent to compete in the Global Workspace.

The LIDA model utilizes two types of special purpose processors—structure building codelets and attention codelets. Structure building codelets build, potentially complex, representational structures in LIDA's CSM. These structures can include, among other things, sensory content from an agent's environment and cued long-term memories (e.g., from Perceptual Associative Memory, Spatial Memory, Transient Episodic Memory, and Declarative Memory). Attention codelets, on the other hand, continually monitor the CSM looking for structures that match their concerns. If found, pre-conscious content and its corresponding attention codelets are formed into *coalitions* that compete for consciousness in LIDA's Global Workspace.

Coalitions consist of attention codelets and the contents for which they advocate. These coalitions are then sent to *compete* within the Global Workspace for conscious "attention." The competition taking place within the Global Workspace module decides to what the system will consciously attend. Whichever coalition has the highest activation has its content broadcast to every LIDA module across the model (i.e., its content is *globally broadcast*). Consciousness consists of, among other things, the frequent serialized broadcast of discrete cognitive moments unfolding across overlapping cycles, that is then typically processed by each module. In other words. Consciousness is discrete and one thing after the other occurs at rapid pace (Baars, 1988). While all of LIDA's modules take in input asynchronously, the serialized nature of the global broadcast facilitates a smooth serialized unfolding of consciousness and, as we shall see, of embodied action. For a general overview of the LIDA model, its modules, and processes, see **Figure 1**.

To be able to address the fact that agents have varying needs, across culture, personal history, and current situations, several variables are attached to structures in the CSM. For example, each structure has an activation value that is used in part to measure its salience. The salience of these structures is used to determine the activation of coalitions containing these structures, modulating their chance of winning the competition for global broadcasting in the Global Workspace. For an in-depth account of salience and motivation in LIDA (see McCall et al., 2020).

[1]For an overview of other cognitive architectures see Kotseruba and Tsotsos (2016).

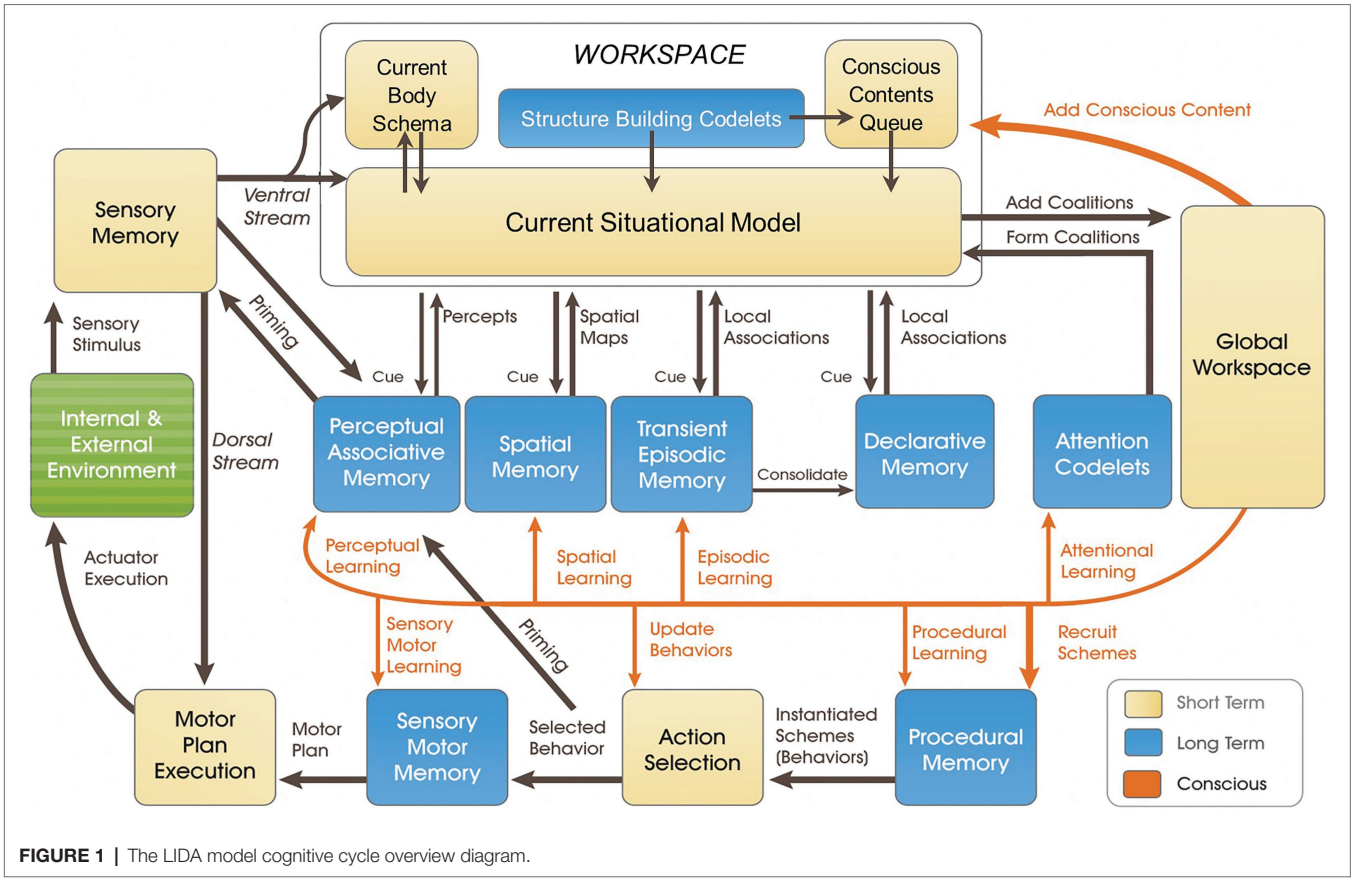



**FIGURE 1** | The LIDA model cognitive cycle overview diagram.

One of the core commitments of the LIDA research program is that the LIDA model is an embodied architecture (Franklin et al., 2013). This means that LIDA agents are biologically inspired in their design, and always in active commerce with their environments. In line with 4E approaches to cognition, LIDA agents are always in the process of answering the question "What do I do next?" Furthermore, constantly answering this question means that all LIDA agents have an "agenda" and in many embodied LIDA agents the agenda stems from the demands of the agent's body.

Debates within embodied cognition often distinguish between weak and strong embodiment (Gallagher, 2011). In rough terms, an approach to cognition is weakly embodied if the body tends to simply be "represented" within a systems central processing. A system is strongly embodied if the arrangement of the systems physical body aids in the constitution of its cognition. However, the LIDA model does not neatly fit into this categorization. The LIDA model uses subsumption architecture (Brooks, 1991), and is in constant sensitive commerce with the environment through its dorsal stream. The LIDA dorsal stream, among other things, directly impact an agent's physical involvement with its world. LIDA agent's also have a body schema that constantly impacts the unfolding of sensorimotor action. At the same time, it is true that the LIDA model also represents its own body within the current situational model. Furthermore, the LIDA cognitive architecture is made so that it can be implemented both in physical and non-physical agents, such as robots or software agents, respectively. Therefore, the LIDA model contains both elements of strong and weak embodiment, and in physical agents, both approaches tend to be in play.

With this overview in hand, we are ready to dig into more detail regarding the LIDA cognitive cycle and action selection. Action selection is of special importance during smooth coping since successful smooth coping requires the skillful selection and execution of the right actions at the right time.

## The Cognitive Cycle

Learning intelligent decision agent's cognitive cycle is divided into an understanding phase, an attention phase, and an action and learning phase (see **Figure 2**). LIDA's cognitive cycle begins with external and internal sensory input, and the construction and updating of structures (i.e., representations) in the Current Situational Model (CSM). Structures that attract the attention of an attention codelet are then brought to the Global Workspace in which they compete for consciousness. The winning structure is broadcast throughout the model, and the system may make a decision to act (internally or externally) through an action selection mechanism. Learning can also occur as the result of each conscious broadcast. While a detailed discussion of learning in LIDA is beyond the scope of this article, it suffices to say that a LIDA agent

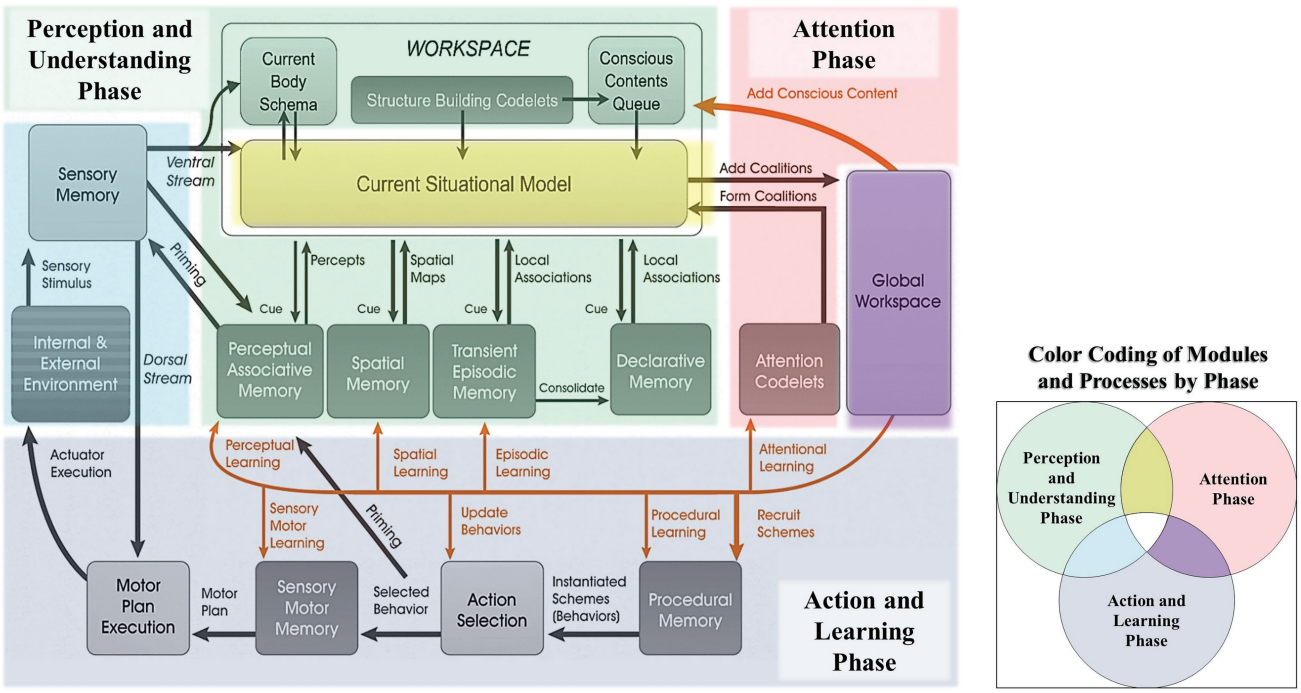



**FIGURE 2 |** The LIDA Cognitive Cycle Diagram color coded. Green modules are involved in the perception and understanding phase, pink modules in the attention phase, and grey modules are involved in the Action and learning phase.

typically learns with each cognitive cycle (as a direct result of its conscious broadcast).

For readers new to LIDA, it is helpful to remember that each cognitive cycle is rapid, lasting only 200–500 ms in humans (Madl et al., 2011), and that LIDA's modules work largely asynchronously and independently of each other. As a result, cognitive cycles can "overlap." For example, the "action and learning phase" from one cognitive cycle can occur concurrently with the "perception and understanding phase" of the next. Thus, while each cognitive cycle is conceptually divided into discrete, serial phases, it is rarely the case that an agent's modules and processes are completely inactive.

## Action Selection

During the action and learning phase of each cognitive cycle, LIDA's Action Selection module will typically select *behaviors* that specify executable (internal or external) actions. This process of action selection is needed for many reasons. For example, it may be the case that many behaviors can accomplish a task, although not all of them equally well. For example, a box might be moved by carrying it, pushing it with one's hands, scooting it with one's foot, or even pushing it with one's head while crawling on all fours. In these cases, Action Selection facilitates the selection of the most situationally relevant and reliable of these behaviors. Furthermore, at any given moment, agents may have multiple, competing desires and goals. Action Selection facilitates the selection of behaviors that are more likely to lead to the most desirable outcomes. Finally, Action Selection coordinates the parallel selection of non-conflicting behaviors. Historically, Action Selection chose one, and only one, behavior at a time. In this paper, we enhance the Action Selection module to include an Automatized Action Selection sub-module (see Section "Smooth Coping in LIDA") that allows for the selection of multiple, non-conflicting behaviors in each action selection event.

Action Selection depends on LIDA's Procedural Memory, a long-term memory module that determinates situationally relevant actions and their expected environmental consequences. In other words, Procedural Memory specifies what actions are available to take, and would happen if they were taken, while Action Selection determines what the agent will do given that knowledge (see **Figure 3**).

As conscious content is globally broadcast throughout all of LIDA's modules, it is received by Procedural Memory, which uses the contents of the conscious broadcast to instantiate[2] *schemes* that are relevant to that conscious content. Instantiated schemes are referred to as behaviors, which are candidates for selection by LIDA's Action Selection module.

Each scheme consists of a *context* (i.e., environmental situation), an *action*, and a *result* (i.e., that action's expected environmental consequences). These can be specified at many different levels of abstraction and generality. Each scheme also contains a *base-level activation*, which serves as an estimate

[2]Instantiation is a specification process. It takes data structures and makes them more concrete. For example, in perception, the "template" for a chair could be instantiated into a specific chair, for example, a chair that is currently in front of an agent.

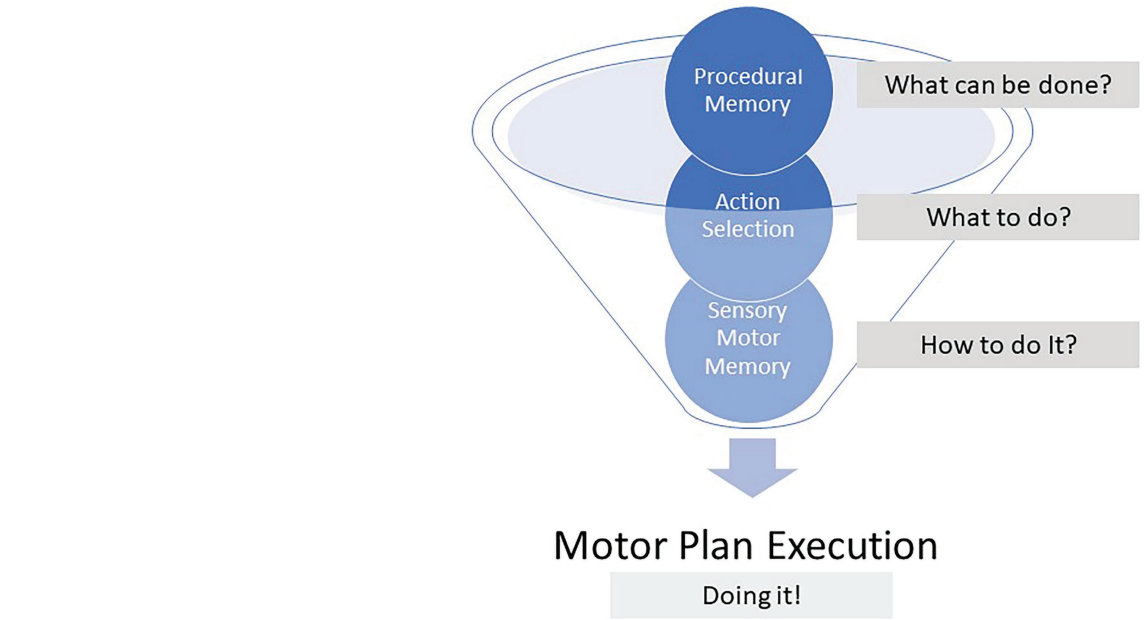



**FIGURE 3 |** To gain a better grasp of the action selection process in LIDA, it is helpful to think of the process as a funneling toward specificity. Procedural memory contains information about things the agent can do under various circumstances at a somewhat abstract level. Action Selection, broadly speaking, chooses "what to do" in the agent's particular circumstance. Sensory Motor Memory decides "how to do it" be picking a motor plan, high specificity, and Motor Plan Execution carries out the motor plan. In this way, actions are procedurally selected with increasing specificity.

of the likelihood that the scheme's result will follow from its action when taken in a given context. For example, a generic "key turning scheme" might specify an action that corresponds to the bodily movements needed to turn a key, the context of being near a lock, and the expected result of that lock being unlocked. Each successful selection and execution of this scheme's action (in the given context) will generally result in an increase in its base-level activation. Similarly, each failure will lead to a decrease in its base-level activation. If, as we might expect, this "key turning scheme" generally succeeds, then it will eventually have a high base-level activation. However, if its context were *underspecified*, for example if it did not limit "key turning" to when an agent is "near a lock," then its action might be taken in inappropriate situations, leading to an unreliable scheme that often fails inexplicably. This unreliability would manifest in the scheme having a low base-level activation.

At this juncture it would be natural to ask, "Wait, is there a scheme for everything? Is there a coffee making scheme? A TV watching scheme? A CrossFit scheme?" First, we must understand that many schemes are culturally specific. A LIDA agent that is implemented in a car factory floor robot does not need a "cool handshake" scheme. However, an agent that exists in a culture in which different handshakes are integral to cultural fluency likely has schemes for different culturally relevant greetings.

Second, we must understand that complex actions are achievable through the execution of multiple simpler actions. For example, riding a bicycle consists of pedaling with both legs, steering, braking, scanning the environment, and much more. Historically in LIDA, the coordination of multiple actions into complex actions has been implemented as *streams of schemes* (see section "Behavior Streams and Skill"). As a result of these streams, LIDA agents do not need to learn unique schemes for every complex action. Rather, seemingly novel complex actions can be manifested through multiple preexisting schemes. In this way, LIDA achieves a form of "transfer learning" (Pan and Yang, 2009). To further facilitate the learning of complex actions, in this paper, we introduce the *hierarchical* organization of schemes (see section "Smooth Coping in LIDA"), which in conjunction with the automatized action selection of actions allows for fluid agential behavior.

When Action Selection chooses a behavior that specifies an *external* action (that is, one intended to modify an agent's external environment), it passes it to LIDA's Sensory Motor Memory for execution. If, on the other hand, the chosen behavior specifies an *internal* action (for example, one used to support mental simulation), it is sent to (or used to spawn) a structure building codelet that updates the Current Situational Model accordingly.

The selection of a behavior can also result in the creation of an *expectation codelet*. Expectation codelets are a type of

attention codelet tasked with monitoring the Current Situational Model for content that matches the expected results of the agent's recently selected behaviors. This temporarily biases an agent's attention toward the environmental consequences of its recent actions, helping to produce a feedback loop between an agent's actions and their results. Thus, in line with enactive and predictive approaches to cognition, action, perception, and prediction are intimately tied together in a feedback loop.

Research on smooth coping generally agrees that smooth coping consists of a series of automatic and consciously controlled actions, as well as both low-level sensorimotor activity and higher-order thought, such as strategizing or monitoring (Christensen et al., 2016; Montero, 2016; Høffding, 2019; Gallagher and Varga, 2020). In other words, smooth coping is a combination of ingrained and automatic processes with conscious and deliberate processes resulting in fluent and skillful action. In LIDA, this is modeled through the combination of four different modes of action selection: consciously mediated action selection, volitional decision making, alarms, and automatized action selection (Franklin et al., 2016, pp. 29–32).

Consciously mediated action selection refers to the many actions an agent performs in which the conscious broadcast is involved, while simultaneously being unaware of the selection processes that go into choosing those actions. For example, in sailing, the sports sailor might be consciously aware of the different ropes on the mast but is *not aware* of the competition in Action Selection that makes her choose the particular rope grip she ends up deploying. Similarly, a tennis player might be consciously aware of the ball as it approaches but is not aware of the action selection process that make him choose the smash over the volley.

Volitional action selection refers to the type of action selection in which the agent is consciously and actively aware of *some* of the selection processes. For example, when an agent is deliberating about what is the best move to make in a board game, and mulling over the different choices, outcomes, and pitfalls, they are doing volitional action selection. By mulling over different possible actions and their outcomes, "options" are created in the Current Situational Model (Franklin et al., 2016). Such options can become conscious and make their way to Procedural Memory, which may then instantiate behaviors based on these options. Action Selection may then choose from among these behaviors. Hence, the first part of volitional action selection is conscious while the second part is unconscious (the conscious broadcast is being utilized but the agent is not aware of the process taking place in Action Selection). In fact, in no mode of action selection is an agent aware of what is happening within the Action Selection module—the module just continuously does its job. In short, during volitional action selection, the agent is aware of the options they are juggling but *not aware* of what is going on "inside" Action Selection.

Alarms are never-conscious processes that bypass the competition in the Global Workspace. If some object or event is recognized by Perceptual Associative Memory as an alarm, the object or event will be sent straight to Procedural Memory to instantiate schemes. Behaviors relevant to alarm content are assigned a high activation value in Action Selection and are typically selected and immediately passed along to Sensory Motor Memory—which in turn passes along motor plans to Motor Plan Execution. Put simply, many agents have experienced acting in an alarming situation, and only becoming aware of their actions after the fact. For example, having a big spider climb on one's arm for a lot of people will result in a series of brushing, jumping, and spasms, in which they are only aware of the threat after the fact. Similarly, in driving, many drivers experience reacting to dangerous situations as fast or faster than they are consciously aware of the situation. Note here that alarms can be both innate as in the spider example or culturally determined as in the driving example.

The final mode of Action Selection is automatized action selection. Automatized actions are overlearned actions where one action can be thought of as calling the next. Selection of automatized actions proceeds unconsciously, that is, selection does not necessarily need content from the conscious broadcast. These are typically the kinds of actions that have been practiced time and time again, and they can be performed without conscious thought. For example, walking on an empty sidewalk is a typical automatized action. It requires little attention, and the agent can simultaneously focus on other matters. In this paper, we go into detail regarding automatized action selection in Section "Smooth Coping in LIDA."

While we go into details regarding automatization in section "Smooth Coping in LIDA" it is worth noting here a core difference between automatized action selection and alarms. Alarm actions revert back to normal functioning once the alarm action has been executed and does not call for further actions. In this way, alarms are a temporary interruption of whatever the agent is doing. Automatized actions on the other hand do not interrupt or take priority over normal processes in the system. Furthermore, automatized actions specify which actions are to proceed them from within the Automatized Action Selection module (more on this in section "Smooth Coping in LIDA").

While in humans this whole process, starting with Procedural Memory, Action Selection, Sensory Motor Memory, and finally Motor Plan Execution, might seem long and laborious, it is important to remember that this process is extremely rapid. Each cognitive cycle typically happens within a few hundred milliseconds (Madl et al., 2011). Thus, when dealing with fast paced dynamic action, as is often the case in smooth coping, the overlapping cognitive cycles are more than sufficiently speedy to make adjustments and act on the fly. Furthermore, we must remember that Motor Plan Execution operates in parallel with all other systems, allowing for non-conscious adjustments to in-flight motor plans. Additionally, the LIDA Sensory Motor System is based on Brooks's subsumption architecture (Brooks, 1991), allowing for rapid agent world interaction.

Similarly, to enactive and predictive processing approaches to mind, LIDA agents are always in the process of adaptively acting; We can say that LIDA agents are perpetually answering the question "What should I do next?" In LIDA, Action Selection continually chooses a behavior among candidate behaviors and sends them to Sensory Motor Memory (unless the action is to deliberate). This ensures that the agent is always in the

process of acting to stay in an optimal adaptive relationship to its environment.

## Behavior Streams and Skill

Smooth coping involves "skill" and "optimal grip." To have an optimal grip on an activity is to skillfully navigate that activity with fluency and ease (Merleau-Ponty, 1945/2012; Rietveld and Kiverstein, 2014; Bruineberg et al., 2021). Concepts, such as "skill" and "fluency," often include being able to execute several actions in an uninterrupted fashion and adjusting those chains of movements to the dynamical real-time changes and demands of the situation (Nakamura and Csikszentmihalyi, 2014).

In LIDA, skill and fluency are, in part, implemented *via behavior streams*. Besides individual schemes, Procedural Memory also contains streams of schemes that can be instantiated. A stream of schemes is a stringed-together series of action schemes that can be collectively instantiated using contents from one or more global broadcasts. The entire instantiated stream of schemes is known as a behavior stream. Once a behavior stream has been sent to Action Selection the module can rapidly select one behavior at a time and pass each of these behaviors on to Sensory Motor Memory (which in turn passes on motor plans to Motor Plan Execution).

For biological agents smooth coping often involves a series of fluent actions. For example, dribbling a basketball, taking three long strides, and then jumping for the slam dunk can occur as one integrated, fluent series of movements. Furthermore, people rarely do just one thing at a time. The action selection process in LIDA, therefore, often involves Action Selection, rapidly picking behaviors from several behavior streams.

Historically, in the LIDA conceptual model, Action Selection has always picked *one*, and only one, action at the time. However, in biological agents, physical actions frequently overlap. Therefore, in this paper we are enhancing LIDA's Action Selection to support the simultaneous selection of multiple actions. Specifically, in addition to the selection of actions one after another by our original action selection algorithm, we are also supporting the simultaneous selection of automatized actions. This is achieved by Action Selection's new Automatized Action Selection sub-module. Developing this sub-module is one of the contributions of this paper.

For example, one can imagine the (haunting) scene of a circus clown riding a unicycle, juggling, and deliberately, maniacally laughing while performatively grinning its teeth. Such a performance requires multiple skilled actions overlapping at once. Even though Action Selection is constrained to choose only one behavior at a time, this does not mean that the *execution* of previously selected behaviors must be sequential. Furthermore, Action Selection can rapidly choose behaviors from multiple concurrent behavior streams, and pass them forward to Sensory Motor Memory for execution.

To be a skilled agent at some activity involves (among other things) having finely tuned, well-rehearsed behavior streams and motor plan templates that can be flexibly adjusted to the demands of the present situation. In LIDA, much of the "skilled" aspects of smooth coping is handled by Action Selection, Sensory Motor Memory, and especially Motor Plan Execution.

As a behavior is sent to Sensory Motor Memory, the system must create a motor plan—a highly concrete plan of bodily movement. Motor plans specify sequences of specific movement commands (the motor commands) that direct each of the agent's specific actuators. Here an actuator simply means one of the physical parts through which an agent acts on the world. For example, a factory robot might only possess a single "arm" actuator. Human beings, on the other hand, have a great many more actuators.

Motor plans and their motor commands react and adapt to rapid incoming data from Sensory Memory through a dorsal stream (Neemeh et al., 2021) to guarantee that the agent's actions are in synch with the most current state of the environment.

Often in smooth coping, an environment may change as an agent is acting on it. For example, being a sports sailor involves skillfully maneuvering the sails of a boat as the vessel is being bumped and rocked by erratic winds and currents. To skillfully complete motor plans during such dynamic situations motor plans constantly react to sensory information through LIDA's dorsal stream as the agent is acting. An agent sailing might issue a motor plan to reach for a specific rope. However, as they are reaching the boat is rocked by a large wave. Instead of continuing the reach in the same fashion, updating the motor plan in real time through the dorsal stream ensures that the agent adjusts their reach, and still successfully grasps the rope.

## Affordances, Action-Oriented Representations, and Behavior Streams

Recent research on smooth coping cashes out much of the skillful interaction loop between agent and environment in terms of affordances and sometimes action-oriented representations (Milikan, 1995; Clark, 2016; Williams, 2018; Gallagher, 2020; Bruineberg et al., 2021; Kronsted, 2021a). Affordances and action-oriented representations are two very similar concepts. Affordances are typically defined as possibilities for actions that exist as a *relation* between an enculturated agent and the environment (Gibson, 1979/2013; Chemero, 2009). Significantly, affordances are ordinarily thought of as a non-representational concept. Action-oriented representations are very similar—but as implied in the name, they are a class of mental representations. Action-oriented representations are representations that also beckon or move the agent into action (Milikan, 1995; Ramsey, 2007; Clark, 2016; Kirchhoff and Kiverstein, 2019).

In LIDA we take a middle-ground approach by using representational affordances. LIDA affordances are conceptualized as representations within the system. For a recent account of how LIDA agents learn and use affordances (see Neemeh et al., 2021). Here it will suffice to say that as LIDA agents become enculturated and trained in various activities, they learn to perceive new affordances upon which they can react. As a LIDA agent gains increased skill, their perceptual system can detect increasingly more fine-grained affordances that can factor into the selection of increasingly fine-grained behavior streams.

There is a careful relationship between action, learning, behavior streams, and affordances. One of the aspects of LIDA that make the model stand out from other cognitive architectures is the "L"—Learning. LIDA agents technically speaking can "learn" something new with every cognitive cycle. With each global broadcast, almost all modules can be updated with content from the broadcast, and each module (including the various memory modules) can perform some function in light of that broadcast. For example, Perceptual Associative Memory might build new connections, Transient Episodic Memory might put together a new event, the Conscious Content Queue adds to the specious present, perhaps Procedural Memory starts building a new scheme, and much more. For a detailed account of learning in LIDA (see Kugele and Franklin, 2021).

In terms of smooth coping, as a LIDA agent acts upon its environment, with each broadcast the agent slowly becomes more familiarized with that environment and the relevant task at hand. Such adaptation includes building more specialized and fine-grained affordances and behavior schemes for those affordances. For example, an agent might not know a thing about Brazilian Jujitsu, but with training, the different movements of opponents become associated with affordances for action or counter action (Kimmel and Rogler, 2018). An opponent going for the rear neck choke—affords putting one's back flat on the mat. An opponent putting their weight in the wrong spot during close guard affords performing a leg triangle choke. There is a virtuous cycle between affordances and their associated behavior schemes. Smooth coping is most often a matter of having fine-grained affordances that make available the use of appropriately fine-grained behavior schemes (see **Figure 4**).

As agents perceives an event, they also perceive the associated affordances. If a coalition containing affordances wins the competition for broadcast in the Global Workspace, then the presence of the affordance in the broadcasted content will help instantiate behavior schemes, and thereby also promote winning the competition in Action Selection.

As mentioned earlier, choosing a behavior (perhaps from a behavior stream) also creates an expectation codelet to facilitate the monitoring of behavior-related outcomes. The creation of expectation codelets not only help bringing action outcomes to consciousness, but also helps ensure that the affordances associated with those action outcomes are also broadcast consciously. Acting on one affordance brings about the next affordance in an action promoting feedback loop. Such a feedback loop is in line with empirical and theoretical literature on affordances that conceptualizes smooth coping as a feedback loop between action and affordances (Di Paolo et al., 2018; Kimmel and Rogler, 2018; de Oliveira et al., 2021; Kimmel and Hristova, 2021; Kronsted, 2021b).

Overall, we see that smooth coping is not a matter of already being skilled at an activity. Rather smooth coping involves the ability to continually improve one's skill and adaptivity. In LIDA, this adaptiveness is built into the flow of information across modules, facilitated by the conscious broadcast.

Of course, smooth coping is not only about knowing "what to do," but also about having sufficiently developed sensorimotor coordination to do so—in layman's terms having the right motor skills. Therefore, the skill cycle in LIDA also includes the agent building and refining increasingly sophisticated motor plan templates. Over many cognitive cycles, Sensory Motor Memory is slowly updated so that the agent is (hopefully) always in a position to know "how to do it" and with a great level of sophistication. Going into detail on how Sensory Motor Memory builds and updates motor plans is outside the scope of this paper. The important takeaway is that LIDA agents consistently update their action capabilities by updating their schemes for "what to do" (behaviors) *and* their plans for "how to do it" (motor plan templates).

Let us take the example of becoming better at sports—in this case, soccer. Through practice, soccer players learn to perceive the field and see it in terms of different opportunities. That is, the player, over time, learns to experience the game in terms of different affordances "in this situation, I can do a long pass, dribble past this guy on the right, or do a short backward pass." Over time, players learn to see the field in terms of affordances that provide possibilities for "what to do" (potential behaviors). However, learning to exploit affordances is also a matter of learning how to concretely utilize the affordance "how to do it" (motor plans). With practice, agents therefore also fine-tune their physical capabilities in part by developing increasingly sophisticated motor plan templates—in the beginning, dribbling and kicking is clumsy, but over time it becomes second nature.

Naturally, doing something as advanced as expert level soccer requires multiple processes—some consciously mediated, others automatic. Hence, next, we will look at how different modes of action selection are interwoven during smooth coping, and the role of automatized action.

## AUTOMATIZATION AND THE AUTOMATIZED ACTION SELECTION SUB-MODULE

One crucial aspect of smooth coping is that it involves both higher-level and lower-level cognitive processes (Christensen et al., 2016; Montero, 2016; Høffding and Satne, 2019; Gallagher and Varga, 2020). Let us return to the clown example. The clown performer who is simultaneously riding a unicycle, juggling, grinning, and talking to select audience members may utilize both consciously mediated, fully conscious, and automatized actions. Thus, to account for such overlapping in action during smooth coping, we need to take a look at how LIDA agents achieve automatization.

An automatized action is implemented as a series of behaviors in a behavior stream that have been mastered to the point in which those behaviors can be selected without mediation from the conscious broadcast—that is automatized behaviors can be selected without the need for sensory input updating. However, the execution of these behaviors may often require sensory input (for example over the dorsal stream or even the conscious broadcast).

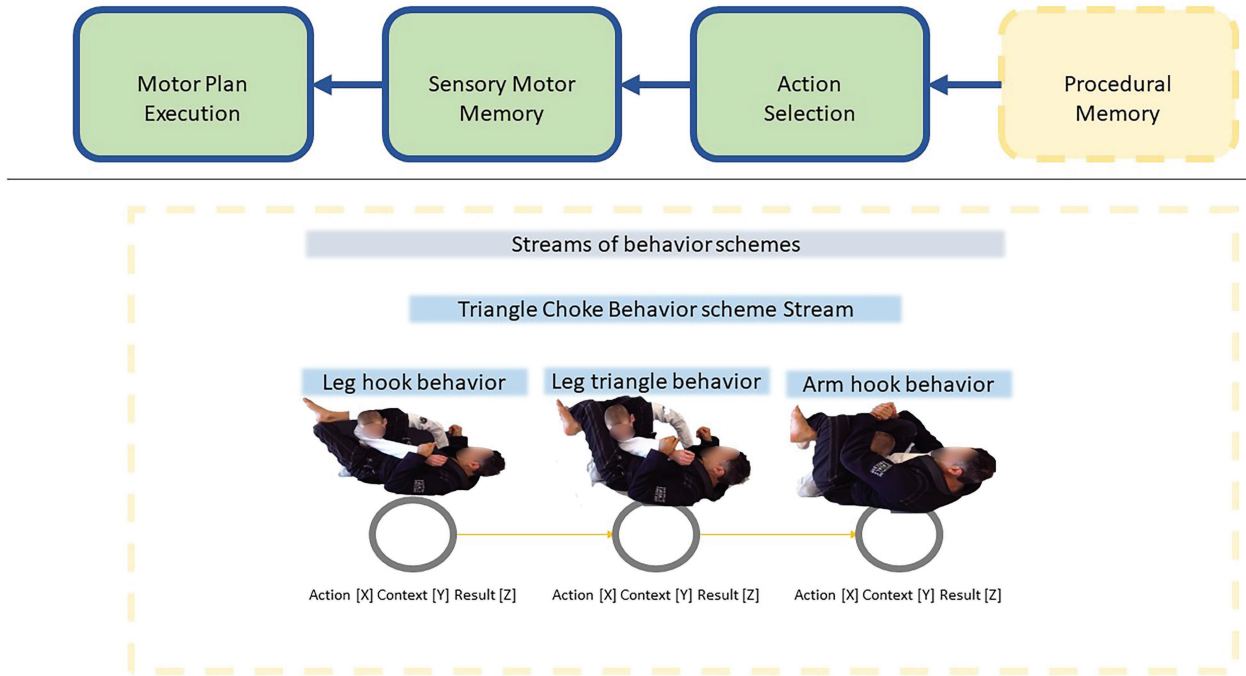



**FIGURE 4 |** Procedural Memory contains streams of specialized behaviors. For example, to perform the Triangle Choke from Brazilian jiu jitsu the agent must first hook their leg around the opponent, form a leg triangle, and then tighten the triangle with legs and arm. These separate behaviors can be executed fluently by having each action linked together in a behavior stream that can have its variables specified with data from the conscious broadcast. By learning actions that are chained together, agents can execute highly specialized behaviors.

For the purposes of smooth coping, it is often important that agents can do several actions simultaneously (for example, pedal and pass, dribble and tackle, punch and block, and the list goes on). In this paper, we therefore introduce a new sub-module to the LIDA model, namely, Action Selection's Automatized Action Selection sub-module (AAS). This sub-module runs in parallel with Action Selection, and repeatedly sends behaviors to Sensory Motor Memory (SMM). For example, in our unicycling clown example, Automatized Action Selection can *repeatedly* choose the automatized behavior "pedal" and send it to SMM.

Having a sub-module that deals entirely with automatized behaviors, and being able to repeatedly select such behaviors, allows for Action Selection to focus in parallel on other forms of action selection, such as consciously mediated action selection or deliberation. Let us return to the example of Jiu Jitsu and the triangle choke. The "triangle choke" is a high-level behavior that consists of several movements (see **Figure 5**): leg hook, triangle hook, arm hook, and the squeeze. When Action Selection selects that high-level behavior, it sends that behavior to the AAS sub-module. From there AAS can select from the component behaviors in the "triangle choke's" behavior stream. In short, Action Selection passes on high-level automatized behaviors to AAS, which then selects from lower-level component behaviors in the high-level behavior's behavior stream. Being able to choose actions in parallel, allows for the Jiu Jitsu practitioner to carefully read their opponent's patterns, and deliberate about what to do next while simultaneously producing complex behaviors, such as the "triangle choke" (**Figures 6, 7**). Smooth coping is often achieved by having Automatized Action Selection working harmoniously in parallel with other forms of action selection.

Automatized Action Selection runs in parallel with Action Selection choosing behaviors from automatized behavior streams (for example, walking, pedaling, dribbling, playing an ingrained song, etc.). Each of the behaviors from the selected behavior stream can be thought of as "calling the next" behavior in that stream. So once a high-level automatized behavior is selected, each of its lower-level behaviors, metaphorically speaking, gets to choose what behavior comes next. For example, if an agent is playing an overlearned piano piece (say *Alley Cat* by Bent Fabric) by way of Automatized Action Selection, each note, which corresponds to a lower-level behavior, "calls the next." Once the first note has been chosen from the "*Alley Cat* Automatized behavior stream," the first note selects the next note upon its completion. This produces the sensation recognized by many musicians as the piece essentially playing itself. This kind of automatization of one action calling the next also ensures that the musician can sing at the same time, lock eyes with the audience, playfully shimmy their shoulders, etc. all at the same time.

In LIDA technical terms, automatized behaviors are "degenerate" behavior streams—they are overlearned actions that *do not include branching options*. The lack of branching options is what allows

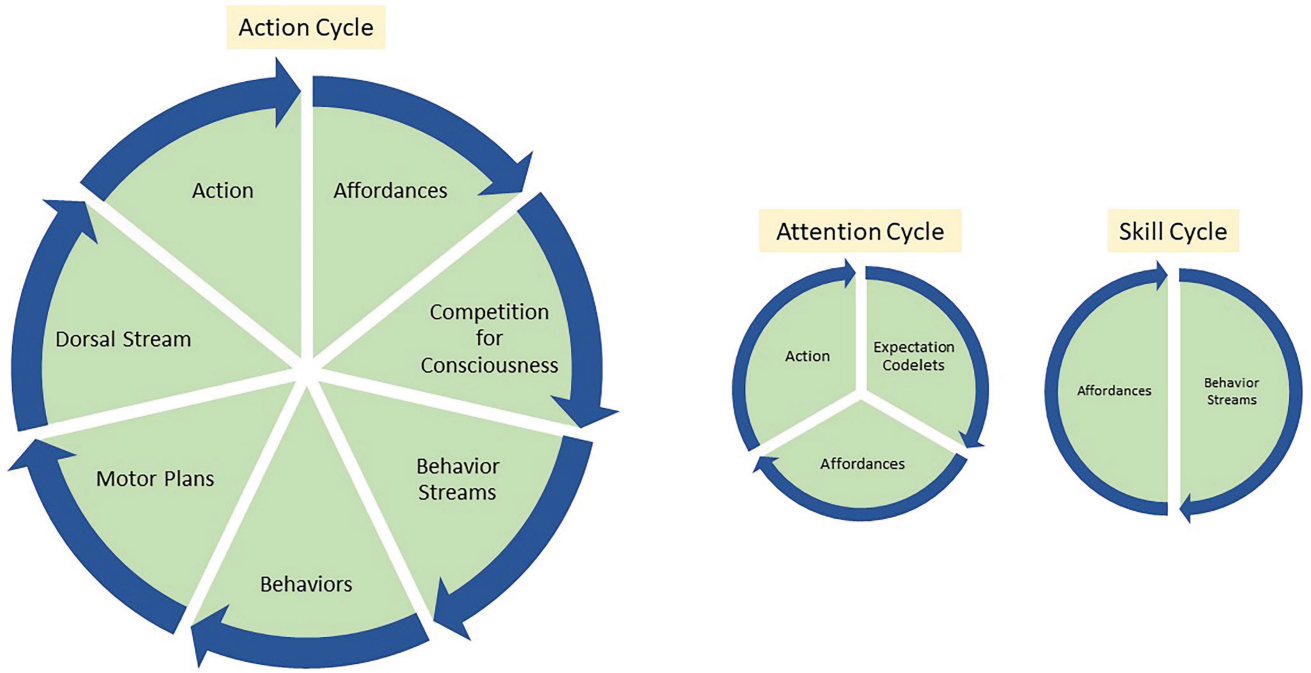



**FIGURE 5 |** Above are three of the virtuous cycles in LIDA agent smooth coping. The first cycle demonstrates the affordance action cycle step by step. The second cycle demonstrates the relationship between expectation codelets new affordances and action. As an agent acts, they also generate expectation codelets and such codelets increases the chance of action-related affordances winning the competition for consciousness. Such biasing of attention in turn creates more actions. Finally, the skill cycle demonstrates how affordances lead to the creation of appropriate behavior schemes and executing behaviors in turn leads to the perception of new affordances.

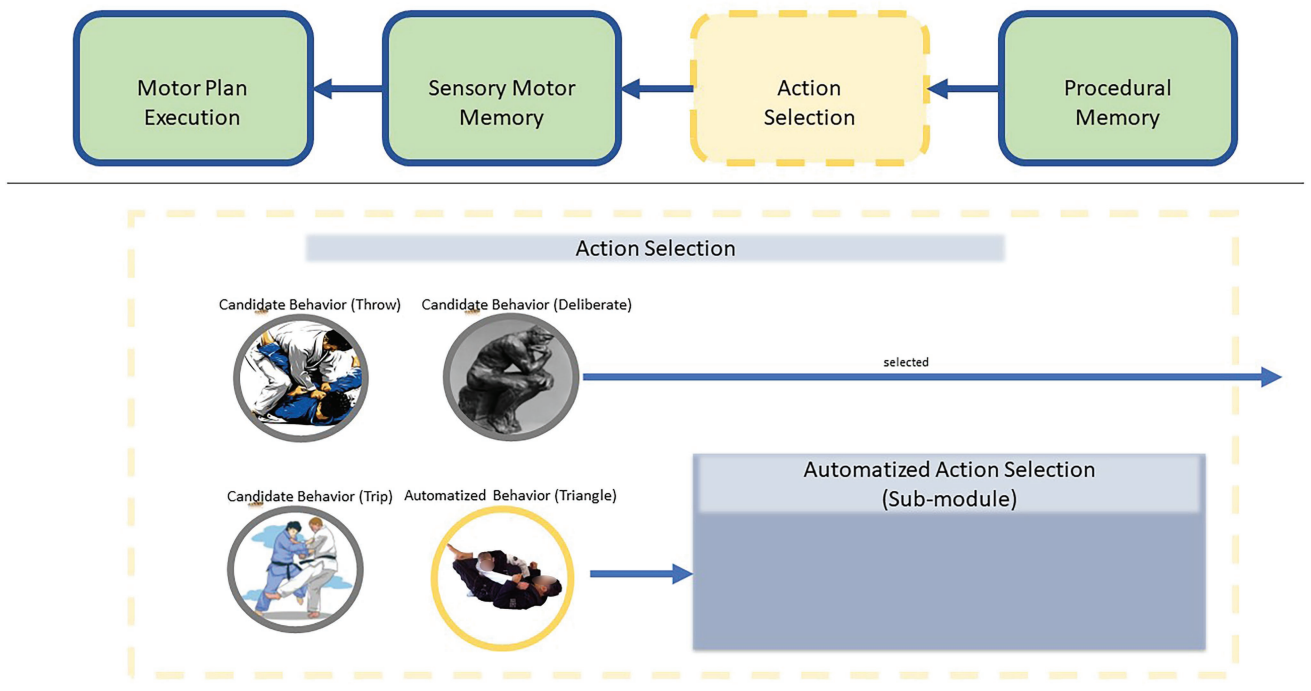



**FIGURE 6 |** Here, we are zooming into Action Selection. In this case, Action Selection is choosing between a wealth of candidate behaviors. In this case, Action Selection chooses the "triangle choke" and passes it on to the Automatized Action Selection sub-module. Action Selection and the Automatized Action Selection sub-module run in parallel to facilitate multitasking. In this case, the agent is choosing to perform a Triangle choke while simultaneously choosing to "deliberate" on what to do next.

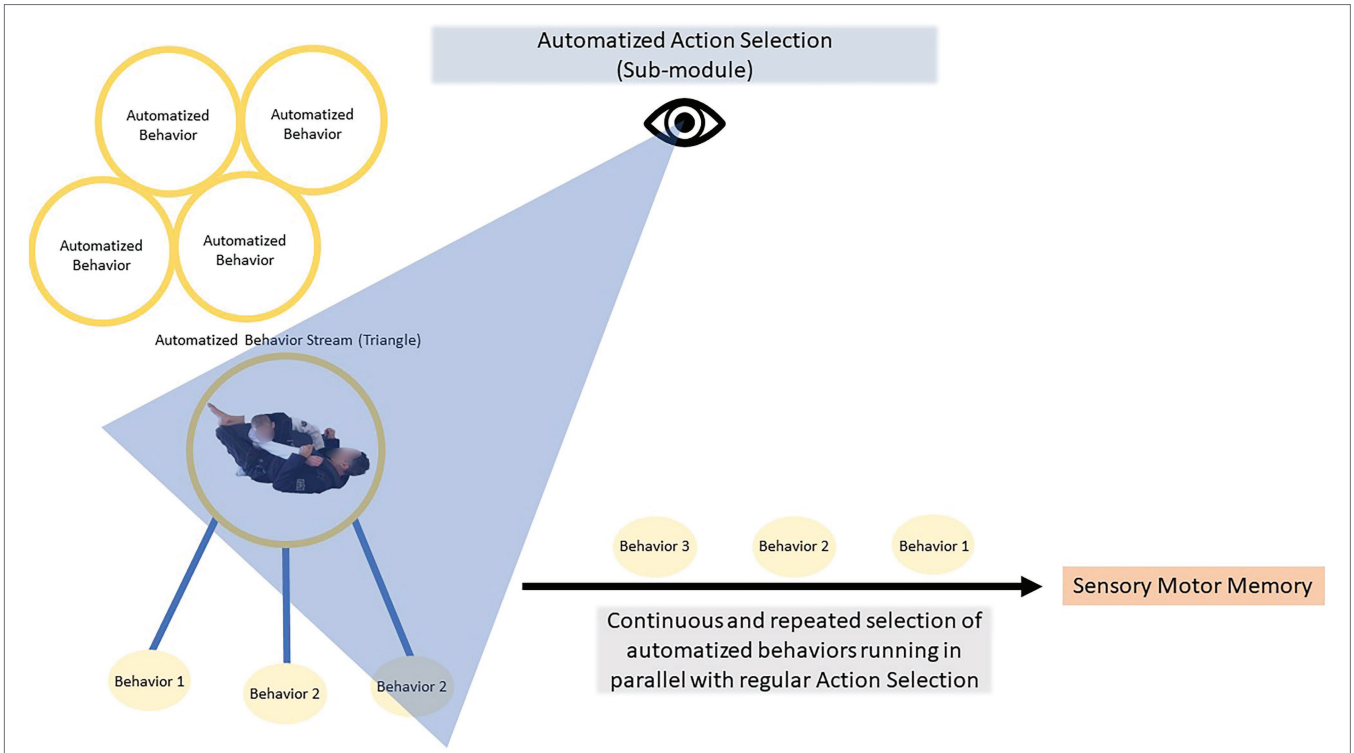



**FIGURE 7 |** The Automatized Action Selection sub-module rapidly chooses one behavior at the time from candidate automatized behaviors (much like regular Action Selection). Like pearls on a string these behaviors are sent forward to Sensory Motor Memory at high speed; all in parallel with whatever might be happening in Action Selection. Differently from regular Action Selection selected automatized behaviors also "calls" for the next action to be selected to insure rapid smooth unfolding of the overlearned series of behaviors.

the behavior to directly "call the next." An automatized high-level behavior for pedaling may contain a behavior for pedaling with the right leg that then calls a behavior for pedal with the left leg—there are no branching options.

Importantly, automatized behavior streams can also be hierarchically structured where each of the behaviors in these streams can correspond to other behavior streams. This capability is critical because the specification of many actions benefits from hierarchical structure, and the reuse of these higher-level behaviors can be more efficient in memory. High-level behaviors often contain multiple behavior streams that must "line-up." For example, to build a Reuben sandwich requires getting bread, mayo, sauerkraut, corned beef, and Swiss cheese, assembling the components, and putting them on a plate. Each of these sub-actions can be automatized and part of its own behavior stream. Collectively, these automatized behaviors contribute to realization of the high-level "Reuben sandwich" behavior.

A deli worker might make and wrap a sandwich like usual without taking the costumer's difficult special order into account "only a little mayo, extra pickles, add sardines!" Making the sandwich differently requires consciously mediated action selection rather than automatization with one action calling the next. This explains why sometimes even when clearly intending to do one thing agents end up doing another because the beginning of the action was of an automatized nature.

It is important to note that although automatized behaviors do not have branching options and call the next action, they still generate expectation codelets. Just as with all other actions in LIDA, the generation of expectation codelets allow the system to keep track of the fulfilment of its actions so that the system may know whether to continue with its behaviors or switch to other behaviors.

As Automatic Action Selection feeds automatized behaviors forward to Sensory Motor Memory, that module can instantiate motor plans that also indicate the "timing" for how long the automatized action needs to be executed for—thereby mitigating the risk of doing something "mindlessly" for too long. In the music example, the motor plans for each note are designated a very short and precise timing. A motor plan for automatized "walking" on the other hand can have the temporal designation "until further notice" within the motor plan. We must remember that while automatization is often good for expert performance, smooth coping involves interwoven types of actions. Relying too much on automatization will often cause the task to fail.

## SMOOTH COPING IN LIDA

One way to describe smooth coping is the use of automatization with intermittent use of consciously mediated actions (see **Figure 8**) as well as other overlapping action selection types toward the

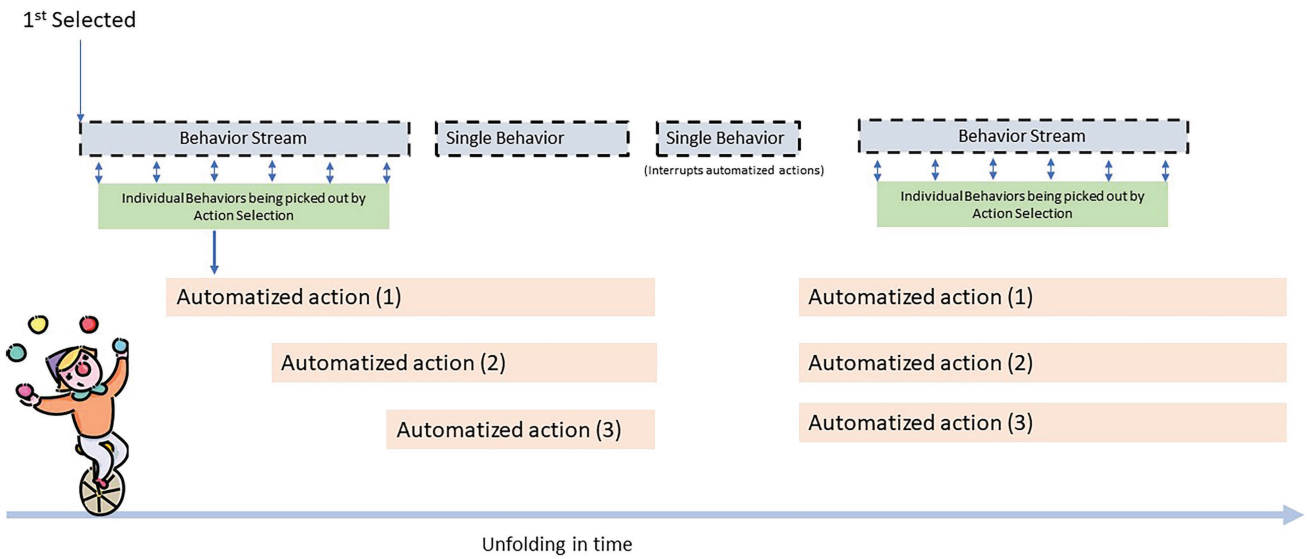



**FIGURE 8 |** Here, we see an example of how an instance of smooth coping could unfold in a LIDA agent. The clown initiates automized actions, such as biking, juggling, and perhaps singing. In this case, the clown starts by biking, then overlays juggling, and finally starts singing (three concurrent automatized behaviors). Intermixed with these automized actions are behaviors picked out from a behavior stream and single behaviors. For example, the clown can turn its head toward select audience members and do a terrifying grin, perhaps do a spin on the bike or in the case of the single behavior that stops all other actions—do a backflip on the bike to then continue the routine.

fulfillment of an intention (Kronsted et al., 2021). The agent is not simply multitasking or simply just doing automatization. Rather, all or most of the agent's cognitive processes are cohering toward fulfilling one intention (completing this difficult recipe, football maneuvers, making it to work through traffic).

If some event forces the agent to abandon the cohering of their actions toward the intention the smooth coping process is interrupted. For example, the unicycling clown is engaging in smooth coping—cycling, juggling, grinning, and singing, all toward the intention of completing their act with a mesmerized audience. However, if a stagehand suddenly runs onto the stage and yells, "You must come at once, your wife is giving birth," then the agent's actions are no longer directed at the distal intention of finishing the act. Smooth coping has been interrupted. Less dramatically, if the phone rings while an agent is cooking, if the agent picks up the phone and attends to the phone call rather than the stove, smooth coping has been temporarily interrupted. The processes can, of course, be re-engaged as soon as the agent puts the phone down. In contrast, if the agent where to continue cooking while talking on the phone the agent can still be said to be smooth coping.

While we have here focused mostly on perception and action selection, and not memory processes, Smooth coping in LIDA is a phenomenon that operates across all modules. As mentioned previously in this paper we here introduce a new addition to the LIDA cognitive architecture—the Automatized Action Selection sub-module. In this section, we briefly go into more detail regarding the different modes of action selection, and then describe their interwoven nature during smooth coping especially in relation to the Automatized Action Selection sub-module. Finally, we provide three concrete case studies to demonstrate how the entire theoretical framework might play out (see section "Conclusion").

## Interwoven Action Selection, and Feedback Loops

We can now see how action selection during smooth coping is achieved in LIDA agents through the interweaving of action selection types—consciously mediated action selection, volitional action selection, alarms, and automatized action selection.

As agents act in a variety of dynamically changing situations, they must deploy different forms of action selection to adaptively achieve their goals. For example, an agent might deploy a series of behaviors and behavior streams to carefully operate a table saw to carve pieces of wood in the right dimensions. Such behaviors and behavior streams might include walking to the table saw, grasping the wood, carefully lining it up on the table, and sliding the wood forward onto the saw while taking aim to ensure a straight-line cut. As the agent is deploying these behavior streams, they might also have intermittent moments of deliberation in which they actively think about which pieces to cut first and how to stack them up in the

right order. The agent might further deliberate about the right dimensions of the cuts, which in turn will trickle down and affect the specifics of the instantiated motor plans and the execution of the actions in Motor Plan Execution.

Since the agent in our example is very skilled at carpentry, they have over years of practice developed automatized behavior streams and highly sophisticated motor plan templates for operating a table saw. So, the agent can operate the saw mostly through Automatized Action Selection.

Perhaps as the agent is working the table saw, their finger gets alarmingly close to the blade, and an alarm is triggered in the system pulling the hand backward. Alarms are importantly a part of the smooth coping flow when they enable the agent to continue with the intended activity. So, in the table saw example, the alarm that stops the agent from cutting off a finger naturally allows for the agent to continue the activity. However, an alarm to shake a large spider off one's hand does not perpetuate the intended activity, and will typically break the smooth coping. The reason to bring up alarms here is to underscore that alarms usually must be learned, and are often skill and context-specific. For example, outside the context of Brazilian Jiu jitsu, getting a nice underhook hug is sweet and comforting. However, within the context of Jiu Jitsu it means the practitioner is about to be swept and likely lose the match. Hence, a context-specific alarm is likely triggered that will make the practitioner pull their arm back and try to close their armpits (to deny the opponent the underhook). Alarms are often an integrated part of mastering a skill since they are rapid and bypass the competition for conscious broadcasting.

Let us return to our table saw example. At some point over years of practice working the table saw has become automatized; the choosing of wood pieces, readying them at the table, and performing the cuts are now done by automatized behavior streams in which one action calls the next. In this way, the agent can repeatedly choose the same reliable behavior streams again and again until the job is done. Automatization allows for the selection of other actions (commonly, consciously mediated or deliberative actions) in parallel with the automatized action unfolding. The worker can operate the table saw (thanks to the Automatized Action Selection sub-module) while yelling at his/her apprentice to correct their form, bring them coffee, or perhaps deliberate about which technique to use for a difficult piece of wood that requires a different technique.

The overarching point is that smooth coping in LIDA involves deploying various forms of action selection each aimed at the task at hand. Be it alarms, consciously mediated actions, deliberative actions, or purely automated actions, each behavior selected coheres toward completing the agent's goal in an adaptive fashion.

At this juncture, we cannot forget that smooth coping involves multiple feedback loops between the agent's actions and changes in the environment. For example, driving behind a car while trying to read a funny bumper sticker on the car, involves having to be at the right range of distances to that car. Too far away and one cannot read the sticker, too close and the cars may collide—the agent must maintain "optimal grip" (Merleau-Ponty, 1945/2012; Dreyfus and Wrathall, 2014; Bruineberg et al., 2021). As already discussed, rapid dorsal stream updating of sensory information in movements updates Motor Plan Execution in action so that the agent can stay in an optimal relationship to their environment during action. There is a constant feedback loop between a LIDA agent's actions and dorsal stream information.

Furthermore, with each action, an expectation codelet is also generated. As mentioned earlier, such codelets scan the Current Situational Model for objects and events related to the expected outcome of the agent's actions. Structures brought to the Global Workspace by expectation codelets are typically highly salient and are very likely to win the competition for conscious broadcast. In this fashion, there is a feedback loop between an agent's actions and their expectations. Through the feedback loop between actions and high activation results, LIDA agents can stay in careful attunement with the unfolding of their activities in dynamic contexts. We see that coinciding with an agent's actions is attention toward the results of those actions which in turn help determine the completion of the intended activity. This is a biasing of attention toward the results of one's actions which in turn helps perpetuate the completion of the intended activity.

Finally, the cognitive cycle in general assists in increasing adaptivity through learning. LIDA agents can update their memory modules with every cognitive cycle (Kugele and Franklin, 2021). In this way, the agent is always slowly but surely moving itself toward a greater degree of adaptivity.

In general, we can think of at least three feedback loops that aid LIDA agents in smooth coping—the general cognitive cycle (adaptivity on a distal time scale), the action attention loop (adaptivity on a proximal time scale), and the action dorsal stream loop (motor adaptivity on a rapid timescale). In short, the cognitive cycle helps with task adaptivity over longer periods of time. Consciously mediated action selection aids in adaptivity in the agent's current context. Automatization, motor plans, and the dorsal stream takes care of rapid in the moment adaptivity (see **Figure 9**).

We have looked at different forms of action selection and how they are interwoven toward the completion of a task during smooth coping. We have also looked at the different feedback loops that comes with these various forms of action selection, and how these feedback loops help the agent adapt to the task across different time scales.

# DISCUSSION

For our discussion, we will apply everything we have looked at so far in three small case studies to see how smooth coping might play out in a LIDA agent in each scenario. We start with the relatively simple example of walking, and move up in complexity to driving, and then short-order cooking.

## Solo Walking

Sam wakes up at 5:00 am to take a daily walk in Shelby Farms Park. The path is a mile loop around a lake, and the early

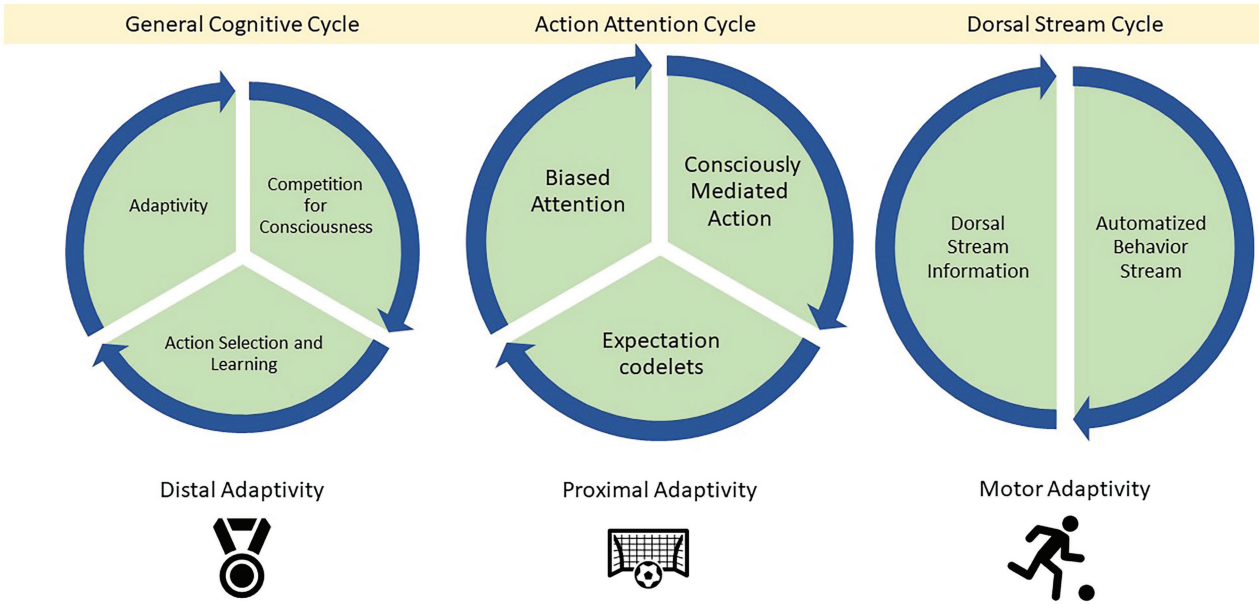



**FIGURE 9 |** Here, we see three feedback loops that aid the agent across different timescales of smooth coping. The cognitive cycle in general aims to keep the agent in an equilibrium with its environment across long time scales. For example, winning a tournament. The attention cycle attunes the agent to their current context and the task(s) they are currently undertaking. For example, the context and task of playing and winning a soccer match. Finally, the dorsal stream cycle aims to keep the agent optimally adapted to their current task at the motoric level across rapid time scales. For example, dribbling, tackling, avoiding other players, shooting at the goal.

hour means that very few others are walking around at the same time.

Sam's system utilizes the automatized behavior stream of walking. As the path curves ever so slightly around the lake, Sensory Memory updates Sam's Motor Plans and motor commands so that Sam adjusts the direction of his body, the height and length of each step and other minor adjustments needed to move through the very accessible flat terrain. Minor differences in the height of the pavement mean that sometimes Sam's Sensory Memory must update his stepping motor commands to be a little longer and a little higher.

Being mostly a matter of automatization, Sam can let his mind wander and think actively about other things in his life that need pondering (should I hop on the Bitcoin craze, is *Squid Game* really that good, what am I doing with my life?). Given that there are no obstacles in the terrain, Sam's systems can simply continue to select and execute automatized walking behaviors. However, no automatized behavior is indefinite, and Sam does still need to periodically check for obstacles. Therefore, Sam still frequently looks at the road ahead and re-selects the automatized walking behavior.

Eventually, Sam notices a pedestrian and their dog approaching. The person and their dog have won the competition for consciousness, and Sam's Action Selection is now choosing between multiple candidate behaviors (while Automatized Action Selection is making sure Sam is still walking). In Action Selection, walking onto the grass or standing still to let the dog and owner pass are the two most salient options. Standing still wins the competition in Action Selection, and Sam lets the person and their dog pass on the narrow path. Choosing this behavior also interrupts the automatized walking behavior.

An expectation codelet is generated looking, among other things, for a clear walking path since this is the expected outcome of Sam's action. While the dog and owner are now behind Sam, the Current Situational Model continues to update. Then the expectation codelet brings the empty path structure to the Global Workspace to compete for broadcasting. Since Sam intends to walk and is expecting to have a clear path, the structure has high activation and may win the competition for consciousness.

As a result of the empty path coming to consciousness, Procedural Memory instantiates relevant schemes including a high-level "walking" behavior. This behavior and its behavior stream are sent to Action Selection. Action Selection chooses the highly relevant automatized "walking" behavior and sends it to the Automatized Action Selection sub-module. As a result, Sam keeps on walking with the Automatized Action Selection sub-module in charge of selecting actions. Now he is again free to continue to think about cryptocurrency, trending TV shows, and existentialism.

## Driving

Sam is done with his existential morning walk. At 8:00 am, Sam drives to work at a local diner. The route is a combination of suburban roads and highway driving, and takes approximately 20 min to complete. Some of the traffic is rush hour traffic.

Sam is utilizing an automatized behavior stream to follow the car in front of him at a safe distance. This of course also

includes the motor plan for safe distance following which is receiving constant dorsal stream updating. Dorsal stream input to the motor plan makes sure that Sam does not push the gas pedal too hard or too softly. Following another car at the appropriate distance in rush hour traffic involves constant adjustment of motor commands to apply the right amount of pressure to the gas pedal.

However, since this is rush hour, Sam also needs to hit the brakes often and at the appropriate pressure. This means that through consciously mediated action selection, the behavior to press the brake is selected and executed at the appropriate level of pressure. Hence, Sam has an automatized car following behavior scheme and motor plan that is being frequently interrupted by the consciously mediated behavior of pushing the brake to remain at the right distance. Each time the brake has been pushed an expectation codelet is generated and helps the resulting distance between cars come to consciousness. The new distance between cars being broadcast in turn helps Action Selection either re-select the automatized follow behavior scheme, or perhaps some other automatized driving behavior.

*Via* consciously mediated action selection Sam decides to activate the behavior stream for changing lanes. Action Selection rapidly chooses each of the behaviors from the lane changing behavior stream. Sensory Motor Memory chooses between motor plans for each of the lane changing behaviors, and Motor Plan Execution begins carrying out the physical movements. In short Sam changes lanes; checks the back mirror, the side mirror, over the shoulder, turns on the blinker, checks again, turns the steering wheel left, turns the steering wheel back to neutral, rechecks windows and mirrors.

Suddenly a person who is texting and driving veers into Sam's lane, and an alarm is triggered. The urgency of the situation means that the closing of the car bypasses the competition for conscious broadcast, and is sent directly to Procedural Memory. Schemes are instantiated and Action Selection chooses an appropriate behavior stream (break and veer). Given the urgency of the situation the break and veer behavior stream has very high salience, and easily wins the competition in Action Selection. Sensory Memory chooses appropriate motor plan templates and instantiates them, and Sam slams the breaks and veers the car away from the reckless driver.

Since an alarm was responsible for the avoidance maneuver, Sam has not yet realized what has just happened. Only approximately 100 milliseconds later, after the event has been recreated in the Current Situational Model, does Sam become "aware" of what just happened. However, during these 100 milliseconds, the break and veering maneuver takes place due to the rapidity of the alarm process. In this way, Sam survives the reckless driver.

During the alarm maneuver expectation, codelets were created, searching the Current Situational Model for the expected results of the dodging maneuver—a safe distance to the incoming driver. As this state of affairs obtains, Sam can now use consciously mediated action selection and choose to aggressively honk at the distracted driver—what a way to start your shift.

## The Short-Order Cook

Sam arrives at work a bit grouchy from the driving encounter. He begins his shift as a short-order cook at a diner. This diner has a counter with the short-order cook behind it and several tables. The diner is particularly busy for the first several hours of the day (people are coming in for brunch and hangover breakfast). Sam is engrossed in work throughout that time and is working on multiple orders simultaneously. The orders are coming in at a fast pace, and many guests are ordering modifications to their dishes (extra cheese, no cheese, chocolate chip pancake on the side, hot sauce on the side, side salad instead of fries, etc.) In addition to making the variety of menu items, several regulars arrive with their special orders and expect to be greeted as they sit down at the counter.

Let us begin with the first order—two eggs benedict, potatoes, and a side of halloumi salad (order one). Upon seeing the order slip, a distal intention is created in the Current Situational Model (finish order one)—this intention cues up information into the CSM regarding halloumi salad, potatoes, and eggs benedict. First, the intention (finish order one) wins the competition for consciousness, and in the next few cycles, structures regarding the current state of the kitchen and structures with information about eggs benedict, potatoes, and halloumi salad, each win a competition for consciousness (given the rapidity of cognitive cycles this is all still within the first second or two!).

At this point, information regarding the state of the kitchen and what to make are now present in the CSM and is broadcast to Procedural Memory. This information is now used to instantiate a multitude of schemes and scheme streams. These candidate behaviors are sent to Action Selection which must now choose "what to do." In this case, the high-level action corresponding to the automatized behavior stream of poaching eggs is selected and sent to AAS. AAS selects behaviors from the "egg poaching" automatized behavior stream and sends them to the Sensory Motor Memory module. Sensory Motor Memory instantiates the chef's highly skilled egg poaching motor plan, and sends it to Motor Plan Execution. This process continues with the other behaviors in the behavior stream being selected by the Automatized Action Selection sub-module where each action can be thought of as calling the next action. Thus, Sam ends up using automaticity to rapidly stir the vinegar–water mix, crack the eggs, and fish them back out.

As Sam is poaching eggs *via* automaticity, a regular customer sits down at the counter (Big Lu). The presence of the regular is highly salient to Sam, and easily wins the competition for consciousness. Procedural Memory upon receiving the global broadcast (containing the content of "Big Lu the regular") instantiates several greeting behaviors, one of which is selected by Action Selection. Simultaneously, the egg poaching automatized behavior is still being executed. In other words, Sam is now stirring the pot rapidly with one hand, cracking eggs into the pot with the other hand, and directing his posture toward the customer while saying, "what's up man, how you been?"

Big Lu tries to greet Sam over the counter with a handshake. But since Sam's hands are full, he needs to use a compensating behavior. The outstretched hand comes to consciousness and instantiates several possible candidate behaviors—one such behavior is to use the elbow to complete the greeting. Choosing this behavior means that a motor plan is instantiated that also takes into account that Sam is still stirring a pot and cracking eggs *via* automaticity. As Sam reaches his elbow over the counter so that Big Lu can high-five his elbow, Sam's motor plans for stirring and egg cracking can be radically adjusted through dorsal stream information and/or through subsequent conscious broadcasts.

As the eggs are being finished, a new order comes in: French toast and scrambled eggs with a side of bacon (order two). This fact comes to consciousness and creates a distal intention for order two which is stored for later retrieval in Sam's Transient Episodic Memory as well as the CSM. Once Sam finishes order one, he can attend to and work on order two. However, at the moment, Sam still needs to assemble order one. The order two intention wins the competition for consciousness, and the intention is broadcast throughout the model, including various short and long-term memory modules (Sam is now working with two distal intentions present in the CSM).

However, Sam is still working on order one. So, Sam is now using consciously mediated actions to carefully assemble the eggs benedict for order one (he needs to grasp and assemble English muffin, ham, poached eggs, and hollandaise sauce).

Given that there are several chefs in the kitchen Sam does not have to make everything from scratch (for example, one worker is at the sauce station, another is at the meats stations). However, Sam does need to know where each component is and the location and activities of his co-workers. This information is updated in Sam's Current Situational Model, including affordances in the environment. For example, if the lid is on the hollandaise pot, the sauce is not available for pouring. However, if the lid is at a tilt, Sam knows from engrained institutional knowledge that his co-worker is done with the sauce. In this case, the pot, therefore, affords "pourability" and Sam uses that information to perform a consciously mediated action of pouring some sauce onto the eggs.

As Sam is assembling the eggs benedict, pouring sauce, and adjusting the garnish, he is comparing the current state of the dish to long-term memory of what eggs benedict generally ought to look like—presentation is half the battle. Furthermore, as he is adding each component to the dish, expectation codelets are continually keeping his attention on track.

Sam puts the finished dish on the service counter for servers to pick up and begins order two, as orders three, four, and five arrive. As Sam is using automatized actions to make more eggs, flipping sauteed potatoes, or stirring, he is also keeping track of each order, and Action Selection is repeatedly sending new behaviors forward. Intermittent with the constant dance between automatized behaviors and consciously mediated behaviors, Sam might need to deliberate. For example, should Sam work on order five instead of four since not all the ingredients for four are ready? An ideomotor process begins with proposers, supporters, and objectors. "No, let us do the dishes in first come first order. That is easiest" "yes, let us put order four on hold to knock down the order we can while we wait for the salmon to finish cooking." Even as Sam is actively deliberating, he is still executing both automatized actions and consciously mediated actions. Ultimately, skipping order four while the salmon is cooking wins the deliberation process, and Action Selection chooses behaviors relevant to making order five.

Around 4 pm the brunch rush is finally over, and Sam gets to hang up his apron and go home. What a day!

## CONCLUSION

Smooth coping is a common phenomenon in high skill activities, such as sports and performance, but also in our daily lives as we navigate the world. Smooth coping generally involves the cohering and centering of cognitive activity toward a task or activity (which is often highly culturally determined).

Learning intelligent decision agent agents engage in smooth coping by interweaving several forms of action selection including; consciously mediated action selection, volitional action selection, alarms, and automatization. Automatizations are overlearned behavior streams that allow for the selection of behaviors without conscious intervention; conceptually for one action to call the next. These automatizations also facilitate the concurrency of automatized action execution. Not only can automatized behavior streams be executed concurrently, but they can also be hierarchically structured. Smooth coping generally involves the biasing of attention and adaptivity toward tasks so that agents can gain an optimal grip on their various contexts. The LIDA model contains various feedback loops across distal, proximal, and rapid timescales that aid the agent in adaptivity. In line with recent embodied and enactive approaches to cognition, LIDA agents are constantly answering the question "what should I do next?" Through interwoven action and perception loops the agent pursues its agenda, and in the process reaches higher degrees of adaptivity across different time scales.

One strength of the smooth coping literature and our exploration of smooth coping in LIDA is that both expert action and quotidian life utilizes the same cognitive resources, and thus we can map a clear progression from novice to expert without the use of any additional "special" cognitive resources. In fact, from the literature on smooth coping and our overview of smooth coping in LIDA we can come to appreciate the complexity that goes into both expert performance and everyday cognition. Despite the ease at which it is performed, smooth coping is an immense achievement for any cognitive system be it artificial or organic.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## REFERENCES


Anderson, M. L. (2014). *After Phrenology. Vol. 547*. Cambridge, MA: MIT Press.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. J. (2019). *ON CONSCIOUSNESS: Science & Subjectivity—Updated Works on Global Workspace Theory*. New York, NY: The Nautilus Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660. doi: 10.1017/S0140525X99002149

Bermúdez, J. P. (2017). Do we reflect while performing skillful actions? Automaticity, control, and the perils of distraction. *Philos. Psychol.* 30, 896–924. doi: 10.1080/09515089.2017.1325457

Bresnahan, A. (2014). Improvisational artistry in live dance performance as embodied and extended agency. *Dance Res. J.* 46, 85–94. doi: 10.1017/S0149767714000035

Brooks, R. A. (1991). Intelligence without representation. *Artif. Intell.* 47, 139–159. doi: 10.1016/0004-3702(91)90053-M

Bruineberg, J., Seifert, L., Rietveld, E., and Kiverstein, J. (2021). Metastable attunement and real-life skilled behavior. *Synthese* 199, 12819–12842. doi: 10.1007/s11229-021-03355-6

Bullard, K., Schroecker, Y., and Chernova, S. (2019). Active learning within constrained environments through imitation of an expert questioner. *arXiv* 1–8. [Preprint]. doi: 10.48550/arXiv.1907.00921

Cappuccio, M. (2019). *Handbook of Embodied Cognition and Sport Psychology*. Cambridge, MA: MIT Press

Cappuccio, M. L., Gray, R., Hill, D. M., Mesagno, C., and Thomas, C. H. (2019). *The Many Threats of Self-Consciousness: Embodied Approaches to Choking under Pressure* (pp. 101–155). Cambridge, MA: MIT Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Christensen, W., Sutton, J., and Bicknell, K. (2019). Memory systems and the control of skilled action. *Philos. Psychol.* 32, 692–718. doi: 10.1080/09515089.2019.1607279

Christensen, W., Sutton, J., and McIlwain, D. J. F. (2016). Cognition in skilled action: meshed control and the varieties of skill experience. *Mind Lang.* 31, 37–66. doi: 10.1111/mila.12094

Clark, A. (2016). *Surfing Uncertainty* Oxford University Press.

Cutsuridis, V., Hussain, A., and Taylor, J. G. (2011). *Perception-Action Cycle: Models, Architectures, and Hardware*. Berlin: Springer Science & Business Media.

de Oliveira, G. S., Raja, V., and Chemero, A. (2021). Radical embodied cognitive science and "Real Cognition". *Synthese* 198, 115–136. doi: 10.1007/s11229-019-02475-4

Dewey, J. (1922). *Human Nature and Conduct*. Baltimore, MD: Holt.

Di Paolo, E., Cuffari, E. C., and Jaegher, H. D. (2018). *Linguistic Bodies*. Cambridge, MA: MIT Press.

Dreyfus, S. E., and Dreyfus, H. L. (1980). *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. Berkeley, CA: California Univ. Berkeley Operations Research Center.

Dreyfus, H. L., and Wrathall, M. A. (2014). *Skillful Coping*. Oxford: Oxford University Press.

Ericsson, K. A., and Kintsch, W. (1995). Long-term working memory. *Psychol. Rev.* 102, 211–245. doi: 10.1037/0033-295X.102.2.211

Fitts, P. M., and Posner, M. I. (1967). *Learning and Skilled Performance*. Belmont: Brooks/Cole Publishing Company.

Franklin, S. (1995). *Artificial Minds. 1st Edn.* Cambridge, MA: MIT Press.

Franklin, S., and Baars, B. (2010). *Two Varieties of Unconscious Processes*. eds. H. A. Collerton and F. LeBeau (Amsterdam: John Benjamin), 91–102.

Franklin, S., and Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. *Intell. Agents III Agent Theories, Archit. Lang.* 1193, 21–35. doi: 10.1007/BFb0013570

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., et al. (2016). A LIDA cognitive model tutorial. *Biol. Inspired Cognit. Archit.* 16, 105–130. doi: 10.1016/j.bica.2016.04.003

Franklin, S., Strain, S., McCall, R., and Baars, B. (2013). Conceptual commitments of the LIDA model of cognition. *J. Artif. General Intell.* 4, 1–22. doi: 10.2478/jagi-2013-0002

Freeman, W. J. (2002). "The limbic action-perception cycle controlling goal-directed animal behavior." in *Proceedings of the 2002 International Joint Conference on Neural Networks*. IJCNN'02 (Cat. No. 02CH37290), Vol. 3, Honolulu, HI, USA. 2249–2254.

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends Cogn. Sci.* 8, 143–145. doi: 10.1016/j.tics.2004.02.004

Gallagher, S. (2011). Faculty of law, Humanities and the Arts - Papers. Available at: https://ro.uow.edu.au/lhapapers/1373 (Accessed September 25, 2021).

Gallagher, S. (2020). *Action and Interaction. 1st Edn.* Oxford: Oxford University Press.

Gallagher, S., and Varga, S. (2020). Meshed architecture of performance as a model of situated cognition. *Front. Psychol.* 11:2140. doi: 10.3389/fpsyg.2020.02140

Ghez, C. (1985). *Voluntary Movement*. eds. E. Kandel and J. Schwartz (Amsterdam: Elsevier Science), 493–501.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Mahwah, NJ: Lawrence Erlbaum Inc. Publishers.

Harnad, S. (1990). The symbol grounding problem. *Physica D Nonlinear Phenomena* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Heidegger, M. (2010). Being and Time (J. Stambaugh & D. J. Schmidt, Trans. Rev. Ed.). New York, NY: State University of New York Press. (Original work published 1928)

Høffding, S. (2019). *A Phenomenology of Musical Absorption*. Berlin: Springer.

Høffding, S., and Satne, G. (2019). Interactive expertise in solo and joint musical performance. *Synthese* 198, 427–445. doi: 10.1007/s11229-019-02339-x

Jonas, H. (2001). *The Phenomenon of Life: Toward a Philosophical Biology*. Evanston, IL: Northwestern University Press.

Kelso, J. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kimmel, M., and Hristova, D. (2021). The micro-genesis of improvisational co-creation. *Creat. Res. J.* 33, 347–375. doi: 10.1080/10400419.2021.1922197

Kimmel, M., and Rogler, C. R. (2018). Affordances in interaction: the case of Aikido. *Ecol. Psychol.* 30, 195–223. doi: 10.1080/10407413.2017.1409589

Kirchhoff, M. D., and Kiverstein, J. (2019). *Extended Consciousness and Predictive Processing. 1st Edn*. London: Routledge.

Kotseruba, I., and Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. *arXiv* [Preprint]. doi: 10.48550/arXiv.1610.08602

Kronsted, C. (2021a). Catching the ghost: house dance and improvisational mastery. *Contemp. Aesthet.* 19, 1–8.

Kronsted, C. (2021b). Using affordances to teach improvisational dance. *J. Dance Educ.* 21, 168–175. doi: 10.1080/15290824.2021.1937184

Kronsted, C., Neemeh, Z. A., Kugele, S., and Franklin, S. (2021). Modeling long-term intentions and narratives in autonomous agents. *J. Artif. Intell. Conscious.* 8, 229–265. doi: 10.1142/S2705078521500107

Kugele, S., and Franklin, S. (2021). Learning in LIDA. *Cogn. Syst. Res.* 66, 176–200. doi: 10.1016/j.cogsys.2020.11.001

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford: Oxford University Press.

Madl, T., Baars, B. J., and Franklin, S. (2011). The timing of the cognitive cycle. *PLoS One* 6:e14803. doi: 10.1371/journal.pone.0014803

McCall, R. J., Franklin, S., Faghihi, U., Snaider, J., and Kugele, S. (2020). Artificial motivation for cognitive software agents. *J. Artif. Gen. Intell.* 11, 38–69. doi: 10.2478/jagi-2020-0002

Merleau-Ponty, M. (2012). Phenomenology of perception (D. A. Landes, Trans.). London: Routledge. (Original work published 1945)

Milikan, R. G. (1995). Pushmi-Pullyu representations. *Philos. Perspect.* 9, 185–200. doi: 10.2307/2214217

Montero, B. G. (2010). Does bodily awareness interfere with highly skilled movement? *Inquiry* 53, 105–122. doi: 10.1080/00201741003612138

Montero, B. G. (2016). *Thought in Action—Expertise and the Conscious Mind. 1st Edn*. Oxford: University Press.

Montero, B. G., Toner, J., and Moran, A. P. (2019) in *Questioning the Breadth of the Attentional Focus Effect*. ed. M. Cappuccio (Cambridge, MA: MIT Press)

Nakamura, J., and Csikszentmihalyi, M. (2014). "The concept of flow," in *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. ed. M. Csikszentmihalyi (Dordrecht: Springer Netherlands), 239–263.

Neemeh, Z. A. (2021). Smooth coping: an embodied, heideggerian approach to dual-process theory. *Adapt. Behav.* 105971232110173. doi: 10.1177/10597123211017337

Neemeh, Z. A., Kronsted, C., Kugele, S., and Franklin, S. (2021). Body schema in autonomous agents. *J. Artif. Intell. Conscious* 8, 113–145. doi: 10.1142/S2705078521500065

Neilson, P. D., and Neilson, M. D. (2005). An overview of adaptive model theory: solving the problems of redundancy, resources, and nonlinear interactions in human movement control. *J. Neural Eng.* 2, S279–S312. doi: 10.1088/1741-2560/2/3/S10

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. New York, NY: W. H. Freeman.

Newell, A. (1973). You can't play 20 questions with nature and win: projective comments on the papers of this symposium. Pittsburgh, Pennsylvania: Carnegie-Mellon University.

Newell, A. (1994). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Pacherie, E., and Mylopoulos, M. (2021). Beyond automaticity: the psychological complexity of skill. *Topoi* 40, 649–662. doi: 10.1007/s11245-020-09715-0

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Papineau, D. (2013). In the zone. *Royal Inst. Philos. Suppl.* 73, 175–196. doi: 10.1017/S1358246113000325

Papineau, D. (2015). Choking and the Yips. *Phenomenol. Cogn. Sci.* 14, 295–308. doi: 10.1007/s11097-014-9383-x

Ramsey, W. (2007). *Representation Reconsidered (1. publ.)*. Cambridge: Cambridge Univ. Press.

Ravn, S. (2020). Investigating dance improvisation: from spontaneity to agency. *Dance Res. J.* 52, 75–87. doi: 10.1017/S0149767720000182

Rietveld, E., and Kiverstein, J. (2014). A rich landscape of affordances. *Ecol. Psychol.* 26, 325–352. doi: 10.1080/10407413.2014.958035

Roseman, I. J., and Smith, C. A. (2001). "Appraisal theory: overview, assumptions, varieties, controversies," in *Appraisal Processes in Emotion: Theory, Methods, and Research*. eds. K. R. Scherer, A. Schorr and T. Johnstone (New York: Oxford University Press), 3–19.

Schack, T. (2004). The cognitive architecture of complex movement. *Int. J. Sport Exercise Psychol.* 2, 403–438. doi: 10.1080/1612197X.2004.9671753

Segundo-Ortin, M., and Heras-Escribano, M. (2021). Neither mindful nor mindless, but minded: habits, ecological psychology, and skilled performance. *Synthese* 199, 10109–10133. doi: 10.1007/s11229-021-03238-w

Sutton, J., McIlwain, D., Christensen, W., and Geeves, A. (2011). Applying intelligence to the reflexes: embodied skills and habits between Dreyfus and Descartes. *J. Br. Soc. Phenomenol. Habit* 42, 78–103. doi: 10.1080/00071773.2011.11006732

Williams, D. (2018). Predictive processing and the representation wars. *Mind. Mach.* 28, 141–172. doi: 10.1007/s11023-017-9441-6

Wilson, M. (2002). Six views of embodied cognition. *Psychon. Bull. Rev.* 9, 625–636. doi: 10.3758/BF03196322

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Kronsted, Kugele, Neemeh, Ryan and Franklin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

HYPOTHESIS AND THEORY
published: 14 April 2022
doi: 10.3389/fnbot.2022.846979

# Situated Neural Representations: Solving the Problems of Content

*Gualtiero Piccinini**

*Department of Philosophy and Center for Neurodynamics, University of Missouri—St. Louis, St. Louis, MO, United States*

Situated approaches to cognition maintain that cognition is embodied, embedded, enactive, and affective (and extended, but that is not relevant here). Situated approaches are often pitched as alternatives to computational and representational approaches, according to which cognition is computation over representations. I argue that, far from being opposites, situatedness and neural representation are more deeply intertwined than anyone suspected. To show this, I introduce a neurocomputational account of cognition that relies on neural representations. I argue not only that this account is compatible with (non-question-begging) situated approaches, but also that it *requires* embodiment, embeddedness, enaction, and affect at its very core. That is, constructing neural representations and their semantic content, and learning computational processes appropriate for their content, requires a tight dynamic interaction between nervous system, body, and environment. Most importantly, I argue that situatedness is needed to give a satisfactory account of neural representation: neurocognitive systems that are embodied, embedded, affective, dynamically interact with their environment, and use feedback from their interaction to shape their own representations and computations (1) can construct neural representations with original semantic content, (2) their neural vehicles and the way they are processed are automatically coordinated with their content, (3) such content is causally efficacious, (4) is determinate enough for the system's purposes, (5) represents the distal stimulus, and (6) can misrepresent. This proposal hints at what is needed to build artifacts with some of the basic cognitive capacities possessed by neurocognitive systems.

**Keywords: neural representation, neural computation, semantic content, situated cognition, embodiment, embeddedness, enactivism, affect**

OPEN ACCESS

***Edited by:***
*Adam Safron,*
*Johns Hopkins Medicine,*
*United States*

***Reviewed by:***
*Roberto Bottini,*
*University of Trento, Italy*
*Guy Dove,*
*University of Louisville, United States*

***Correspondence:**
*Gualtiero Piccinini*
*piccininig@umsl.edu*

***Received:*** *31 December 2021*
***Accepted:*** *23 March 2022*
***Published:*** *14 April 2022*

***Citation:***
*Piccinini G (2022) Situated Neural Representations: Solving the Problems of Content. Front. Neurorobot. 16:846979. doi: 10.3389/fnbot.2022.846979*

## THE PROBLEMS OF CONTENT

Explaining cognition in terms of neural computations over neural representations, as mainstream cognitive neuroscience does, raises tough foundational questions. Among the most difficult are a cluster of related problems pertaining to the putative semantic content of neural representations. I will refer to them as the *problems of content*:

1. *The source of original content* (cf. Haugeland, 1998; Jacob, 2019). The semantic content of public language and other public symbolic systems is *derivative*—that is, it seems to derive from other entities, the symbols' users, whose states appear to possess semantic content of their own. For instance, the word "burro" means *butter* in Italian but *donkey* in Spanish; the very same physical symbol—the same sequence of phonemes or letters—can mean different things in different languages. The most plausible explanation is that the content of words such as "burro" derives from the speakers of the different languages within which the words occur. In contrast, if intelligent agents operate via representations internal to their neurocognitive systems, the semantic content of their internal representations must be *original*—it cannot be derived from other semantically contentful sources on pain of vicious regress. But there is no consensus on how neurocognitive states can acquire original semantic content.
2. *The coordination between vehicles and their content* (cf. Fodor, 1994, p. 12ff, 86; Piccinini, 2004, p. 405). Representational explanation requires that vehicles be processed computationally in a way that matches their content. For example, suppose we want a computer to perform inferences about animals. The inferences the computer performs must match the meaning of its various symbols: for instances, from "there is a dog" the computer may infer "there is a barking animal" but may not infer "there is a meowing animal"; the opposite must hold for "there is a cat." In ordinary artificial computers, the match between computations and the semantic content of the vehicles is accomplished by the programmer, who can independently access both the computational vehicles and their content and program the computer accordingly. In the case of neurocognitive systems, however, there is no external programmer. Thus, it is unclear how computational vehicles and the computations performed over them can be matched to appropriate semantic contents. It seems that any putative mechanism tasked with matching vehicles and the way they are processed to the vehicles' semantic content must have independent access to both vehicles and their contents, so that it can match them accordingly. This would require that vehicles and contents be accessible independently of one another within the neurocognitive system, which does not seem possible.
3. *The causal efficacy of content* (cf. Stich, 1983; Dretske, 1988; Fodor, 1994). Insofar as representations explain behavior, they appear to do so in virtue of their content. For instance, suppose that my dog Cinnamon licks my face because she is happy that I'm back home, and this is cashed out in part in terms of Cinnamon's neural representation whose semantic content is that *I'm back home*. Such semantic content is supposed to contribute to the explanatory power of representations. For, if Cinnamon's representation, causing her to lick my face, had a different content—e.g., *that the cat is meowing*—then a representational explanation of why Cinnamon is licking my face would fail. But what causes behavior is the vehicle that carries the content, which is what the system physically processes. Since the causal work is done by the vehicle, the semantic content has no causal work left to perform. In addition, semantic content appears to be relational in a way that undermines its causal efficacy. For semantic content is a relation between the vehicle and what it represents, and that does not seem to be the sort of thing that can play a causal role. If these observations are correct, then semantic content plays no causal role. If so, content is epiphenomenal and representational explanation is illusory. The genuine explanation of behavior is causal and, therefore, it can't appeal to semantic content.
4. *The determinacy of content* (cf. Shea, 2018; Neander and Schulte, 2021). It seems to many that a notion of representation worthy of its name should come with determinate semantic content—the kind that can be expressed by a proposition and evaluated as true or false or, in the case of concept-like representations, the kind that can be expressed by a linguistic predicate. But theorists disagree about what content neural representations have. A classic example is what the frog's eye tells the frog's brain (Lettvin et al., 1959). Even theorists who agree pretty closely on what determines the semantic content of neural representations have offered different interpretations of the internal signals that allow frogs to detect, catch, and eat bugs. They have proposed that the signals' content is (i) *fly (there now)*, (ii) *something small, dark, and moving (there now)*, or (iii) *food (there now)*. There is no consensus on how to resolve this disagreement. This suggests that putative neural representations lack determinate contents after all, which in turn suggests that neural vehicles are not representations properly so called.
5. *The distality of content* (e.g., Dretske, 1988; Shea, 2018; Neander and Schulte, 2021). Between a stimulus and a neural state, there is a causal chain involving many intermediate causes, all of which correlate with the internal state and all of which may be said to cause the internal state. For instance, a visual stimulus such as a flower in a garden causes patterns of light waves that travel through the air, which cause activation patterns in the retinas, which cause spike trains to travel through the optic nerve, which cause activation patterns in the lateral geniculate nucleus of the thalamus, etc. Many naturalistic theories of content assign content at least in part based on the relation between a representation and what causes it (Adams and Aizawa, 2021). If the content of a representation is determined by what causes it, however, it's unclear why a neural state should represent the distal cause—e.g., the flower—rather than any of its more proximal causes.
6. *The possibility of misrepresentation* (e.g., Dretske, 1986; Fodor, 1994; Neander and Schulte, 2021). If a system can represent, it should also be able to *mis*represent. For instance, if visibility is poor, a system might mistake a horse for a cow, thus representing a horse as a cow. As noted above, however, many naturalistic theories of content appeal to the relation between a representation and what causes it. Accordingly, if a representation is caused by a horse, its representational content should be *horse*, not *cow*. But then it's unclear how a representation can ever misrepresent. There is no consensus about how a naturalistic theory of content can account for the possibility of misrepresentation.

The difficulty of one or more of these problems has led many theorists, including many theorists of situated cognition, to reject neural computations, neural representations, or both (e.g., Casper and Artese, 2020). According to such theorists of situated cognition, cognition is situated—that is, embodied, embedded, enactive, and affective—*as opposed to* representational and computational (e.g., Thompson, 2007). Many others continue to maintain that cognition involves computation over representation, and some have correctly pointed out that cognition being representational and computational is compatible with cognition being situated (e.g., Clark, 1997; Miłkowski, 2017). Even among such compatibilists, however, there is no consensus on how to fully solve the problems of content.

I will argue not only that situatedness is compatible with computation and representation, but also that the situatedness of neurocognitive systems and, as a consequence, *the situatedness of neural computations and representations is the very key to solving all the problems of content at once*. Specifically, I will argue that neurocognitive systems that are embodied, embedded, affective, dynamically interact with their environment, and use feedback from their interaction to shape their own representations and computations (1) can construct neural representations with original semantic content, (2) their neural vehicles and the way they are processed are automatically coordinated with their content, (3) such content is causally efficacious, (4) is determinate enough for the system's needs, (5) represents the distal stimulus, and (6) can misrepresent.

Caveat 1: The most successful account of the semantic content of internal representations is informational teleosemantics (Dretske, 1988; Neander, 2017; Shea, 2018). Roughly, according to informational teleosemantics, the semantic content of (indicative) representations is the information they have the function to carry. Existing versions of informational teleosemantics go at least part of the way toward solving problems 1 and 4-6. This is largely because teleosemantics already includes an important element of situatedness: the teleofunctions that give teleosemantics its name are *wide* functions—functions that reach into the organism's environment. That said, problems 2 and especially 3 are harder to crack; I will argue that solving them along with fully solving the others requires a more thorough appeal to the organism's situatedness. As I will point out, the recent literature contains hints that the solutions to the problems of content are to be found in the situatedness of neural representations. The considerations to follow are intended to (i) improve on existing versions of teleosemantics by (ii) making points that are either overlooked or only implicit in the teleosemantics literature, thereby (iii) showing how the situatedness of neurocognitive systems contributes to solving the problems of content and (iv) providing a unified solution to the problems of content.

Caveat 2: I will not propose a complete account of intentionality. For present purposes, intentionality is the property of indicative mental states, such as beliefs, to the effect that they can be attributed a propositional content with full-blown truth conditions, as opposed to the kind of accuracy conditions that I will adopt as a standard for the kind of (nonpropositional) neural representations that make up the bulk of the cognitive economy of most animal species. Explaining intentionality involves explaining fully determined propositional contents, referential opacity, representation of nonexistent objects, and other phenomena that go beyond the scope of this essay. What I will do is propose a solution to some of the most difficult problems faced by an account of basic neural representations with original semantic content, problems which lie at the foundation of any naturalistic theory of intentionality. Fully accounting for intentionality itself is a separate project, which will require additional work (for steps in that direction, see Morgan and Piccinini, 2018; Piccinini, 2020b).

Caveat 3: I will set phenomenal consciousness aside. The relationship between neural representation (and computation) and phenomenal consciousness is challenging territory that lies outside the scope of this paper (for more detailed discussion of options and some hints at the direction that appears most promising, see Piccinini, 2020a, Ch. 14 and Anderson and Piccinini, unpublished, Ch. 7).

Caveat 4: In addition to embodiment, embeddedness, enaction, and affect, situated approaches also include the thesis of extended cognition, that is, that some cognitive states or processes occur outside the skull. Whether cognition is extended does not affect my argument, so I will remain neutral about that.

## BASIC FRAMEWORK

I will adopt a theoretical framework defended in detail by Piccinini (2020a, 2022). Here I will briefly recap the main aspects that are relevant to this project. This section is intended primarily for philosophers; non-philosophers can skip it without too much loss.

The universe consists of many objects that stand in compositional relations: small objects compose larger objects, which compose ever larger objects until all objects, taken together, compose the whole universe. Objects have natural properties, including relational properties. There are three types of property: qualities, such as shape and size; causal powers, such as the ability to fire action potentials; and structural properties, such as being made of neurons and glial cells arranged in a certain way. An object's properties are invariant aspects of the properties of that object's parts. A composite object itself is an invariant under certain transformations in its parts.

The objects we are concerned with are substantive wholes, namely, objects whose (proper) parts change their properties when they come to stand in organizational relations such that the parts compose such wholes. For instance, pluralities of disconnected neurons and glial cells cannot perform nontrivial cognitive functions. When they are connected together and sustained by an organism's metabolism, however, neurons and glial cells form nervous systems, thereby acquiring the ability to send signal to one another and, collectively, to perform nontrivial cognitive functions.

It's important to note that causal powers require disposition partners for their manifestation and are typically individuated by the manifestations they have when they encounter their

partners (Martin, 2008). For instance, the very notion of a *signal* presupposes that the signal is sent to one or more receivers. Accordingly, the power to send a neuronal signal presupposes a communication channel through which a neuron sends the signal to one or more receivers. Thus, causal powers include an intrinsic aspect—what the object can contribute to a manifestation—and a relational aspect—what the object must be related to in order for the manifestation to occur.

It's also important to note that a property can be both the manifestation of an object's causal powers as well as a causal power of its own. For instance, a truck's momentum is both a manifestation of its power to set itself in motion and a causal power of its own, which can be transferred to other objects in case of collision. Later I will argue that semantic content is a case of this sort: both the manifestation of some of a neurocognitive system's causal powers and a causal power of its own.

Some systems contain *mechanisms*. For present purposes, mechanisms are subsystems composed of different types of part, each with its own specialized powers, and the parts are organized in such ways that each part meets disposition partners in some of the mechanism's other parts and portions of the environment of the system. As a result, mechanisms have powers that their parts, when they are not organized to form the mechanism, could never have.

Some special mechanistic systems are *organisms*. What counts as an organism is a difficult question that I cannot address in depth here. Suffice it to say that organisms have special closure properties such that their parts are mutually involved in maintaining the organization of the system[1]. Organisms include sets of entities each of which can be produced from other entities within the set (Kauffman, 1993), organisms exert work to maintain internal constraints that in turn are necessary to produce the work (Kauffman, 2002), their processes are mutually constrained in such a way that each constraint is generated by at least one other constraint (Montévil and Mossio, 2015), and their behavior, broadly construed to include metabolism, must result at least sometimes in a mutually supportive set of conditions that include survival, development, reproduction, and helping others (Piccinini, 2020a, p. 68). I call the latter four conditions *goals* in the following minimal sense: they require work and, if all members of a population fail to fulfill them, eventually the population goes extinct. Thus, for organisms to continue to exist, the four goals must be pursued and fulfilled at least sometimes by some organisms.

Since organisms have goals that they must pursue, their traits (parts and their properties) as well as the artifacts they build and use may contribute to such goals. Contributing to such goals in a stable way is what I call the biological *function(s)* of such traits and artifacts. Token traits and artifacts that belong to a type some of whose tokens are able to perform a function may be said to have that function even though they cannot perform it or cannot perform it at the appropriate rate in appropriate situations. Thus, this is a normative notion of function: traits and artifacts can function incorrectly, malfunction, or completely fail to perform their function[2].

Some organisms have specialized control organs—namely, nervous systems—whose function is to direct the behavior of the organism as a whole in response to environmental, physiological, and developmental conditions. Fulfilling control functions requires transducing different kinds of external signals into internal vehicles that allow the control organ to integrate different sources of information, build and update internal models of the body and environment, and use such models to guide and direct behavior. Since the function of the vehicle is to encode different sources of information as well as guide the control of a complex organism, the vehicles themselves are defined in terms of such functions, not any particular ways in which the vehicles are physically implemented. I call such vehicles medium-independent, and the manipulation of such vehicles in a rule-governed way, which is needed to perform control functions, *computation* in a generic sense. While neural processes are computational in a generic sense, there are good reasons to conclude that they are sui generis computations—neither digital nor analog[3].

## NEURAL STRUCTURAL REPRESENTATION

There is a widespread consensus that the notion of representation that is relevant to cognitive neuroscience is that of *structural* representation[4]. To a first approximation, a structural representation is a model of a target that can guide behavior with respect to its target. For example, a map of a territory is a structural representation. More precisely, I define a structural representation as a system that has the function of possessing the following four features: (i) a partial isomorphism (homomorphism[5]) to its target, (ii) being activated by signals coming from its target, (iii) the ability to guide

[1]This self-organizing feature of organisms has long been emphasized by what is sometimes called *autopoietic enactivism* (Varela et al., 1974; Ward et al., 2017).

[2]Some complex organisms have sentience and sapience, which give rise to nonbiological functions, which are stable contributions to nonbiological goals. Nonbiological functions are not especially relevant here.

[3]Roughly, digital computations can operate over sequences of discrete states, analog computations can operate over continuous variables, and neural computations operate over spike trains; within spike trains, spikes are distinct from one another, which makes them somewhat similar to discrete states, but their frequency and sometimes their timing are functionally significant, which makes them somewhat similar to continuous variables. Since the vehicles of neural computation shares similarities and differences with the vehicles of both digital and analog computation, neural computation is sui generis. A more detailed treatment is in Piccinini (2020a, especially Chs. 6 and 13).

[4]The following account of neural representation and its content was influenced most directly by Piccinini (2020a, Ch. 12), which is a descendant of Thomson and Piccinini (2018), and by Lee (2021). Other important recent sources that influenced me on the problems of content and related matters include Gładziejewski (2015), Ramsey (2016), Gładziejewski and Miłkowski (2017), Miłkowski (2017), Neander (2017), Buckner (2018, forthcoming), Dewhurst and Villalobos (2018), Lee (2018), Shea (2018), Millikan (2021), Poldrack (2021), and Bielecka and Miłkowski (2020).

[5]Some authors prefer the notion of similarity to that of homomorphism. For a recent account of semantic information carried by structural representations in terms of similarity, see Miłkowski 2021.

behavior with respect to its target, and (iv) the ability to be decoupled from signals coming from its target (and therefore to guide behavior with respect to its target even when its target is not directly activating the representation)[6].

Ramsey argues that, in addition to defining structural representation in such functional terms (a model that can guide behavior), we also need an account of the semantic content of structural representation (Ramsey, 2007, 2016). He points out that many theorists either fail to distinguish between the functional role of structural representations and their semantic content or they simply ignore the functional role.[7] Ramsey concludes that, in addition to an account of functional role along the lines I gave above, we also need an account of the representations' semantic content. The most successful account of the semantic content of structural representations is informational teleosemantics, which says, roughly, that the semantic content of a structural representation is the information it has the function of carrying about its target (Dretske, 1988; Neander, 2017; Shea, 2018). For present purposes, that a state carries information about a target means that the occurrence of that state raises the probability that the target is also occurring.

I agree that the notion of structural representation is the relevant one, and I will endorse a version of informational teleosemantics. I add that neural representations have special features such that, when the relevant notion of structural representation and the relevant teleosemantic theory are formulated properly, *the vehicles of neural representations and their semantic content are two sides of the same coin*. That is, the same functional properties that turn a system of internal states into a neural representational system are also sufficient to give such internal states their semantic content[8]. I will also argue that, once we gain an adequate account of the ontology of original semantic content, the content of neural representations is an aspect of their causal powers—the power to track their target and, as a consequence, to guide behavior with respect to their target.

For present purposes, a neural structural representation is a state of a *simulation* of a target, where a simulation is a system of states, homomorphic to their target, which can evolve to match the evolution of their target to some degree of approximation. In addition, a *neural* structural representation is a state of a system whose functions includes the following:

1. To build and maintain a simulation of its body and environment.
2. To use the simulation to guide behavior by issuing motor commands.
3. To use information from the body and environment together with its own motor commands to update the simulation.

A system that performs the above functions has all the four features of structural representations. By definition, the simulation it builds and maintains is homomorphic to its target and can guide behavior. By relying on information from the body and environment to update its internal states, the system gets activated by signals from its target. Finally, since the simulation is a dynamical model that can evolve on its own in a way that can match its target, its states can be decoupled from their target.

A system that performs the above functions already has all that's needed for its states to have semantic content according to informational teleosemantics[9]. This is because, since one of the system's functions is building a simulation of its environment and updating it using information from the environment, the states of the simulation carry information about environmental states. We may conclude that one of the states' functions is tracking their targets, or we may prefer to say that they track their target, when they do, due to the function of the system as a whole; regardless, this is enough for a viable teleosemantics. It is in virtue of the information they carry about their targets that such states can guide behavior with respect to their targets.

Now let's consider the metaphysics of the semantic content of this kind of structural representation. Recall from the previous section that causal powers include an intrinsic aspect—what the object can contribute to a manifestation—as well as a relational aspect—what the object must be related to in order for the manifestation to occur. Each state of the sort of simulation we are discussing has an intrinsic aspect—the ability to receive, process, and send signals—and a relational aspect—the relations to the rest of the system. It is the relations to the rest of the system, which in turn is related in appropriate ways to the body and environment, which enable each internal state to receive and send signals carrying information about their target and to guide behavior on that basis.

On one hand, the system has learned to activate each internal state to track specific targets and predict the target's evolution. Thus, when the system functions correctly, each internal state sends its signals under appropriate circumstances (information is flowing in either directly from the target or from other internal states that carry information about the target, including past states of the system). On the other hand, when the system functions correctly, each internal signal can be used to guide

[6]The notion of representation primarily under discussion here is that of *indicative* representation, whose function is to fit the world. There are also *imperative* representations, whose function is to change the world to fit them. I discuss imperative representations in Piccinini (2020a, Ch. 12).

[7]Facchin 2021 has recently questioned whether structural representations' functional role is genuinely representational. Roughly, he argues that (a) receptors are not genuinely representational, yet (b) some receptors are structural representations; therefore, (c) some structural representations are not genuinely representational. Even as Facchin presents his argument, it leaves room for *some* structural representations to be genuinely representational, which is all I need. In any case, Facchin does not establish (a); instead, he makes a plausible case that (a') *some* receptors are not genuinely representational. Needless to say, (c) doesn't follow from (a') and (b).

[8]For an account that goes somewhat in the same direction, see Shea (2018, p. 10, Chs. 3 and 4).

[9]My version of teleosemantics is not based on the usual, selectionist account of functions, according to which functions are selected effects (e.g., Neander, 2017); it is based on the goal-contribution account of function I briefly reviewed in the previous section according to which functions are stable contributions to the goals of organisms. One advantage of this innovation is that it makes it possible for semantic content to be causally efficacious.

behavior in the relevant way—i.e., with respect to its target. As a result of the combination of its intrinsic and relational properties, each internal state has the causal power to track its target, predict the target's evolution, and guide behavior with respect to its target (to the extent that the system is performing its representational function)[10].

The semantic content of a neural representation is a manifestation of its power to track its target and predict its evolution. It is also the causal power to guide behavior with regards to its target. Content is often represented by using that-clauses; for example, "the cat is on the mat" means *that the cat is on the mat*. This is an inadequate way of expressing the content of neural representations, at least in the general case. On one hand, typical neural representations do not enter the kind of explicit inferential relations that linguistic representations, whose content is expressed by that-clauses, can enter. In addition, their correctness conditions are a matter not of truth or falsehood but of degrees of accuracy with which a target is tracked. On the other hand, however, neural representations are rich in detail, connected to other representations, dynamical, and predictive of their target's evolution in a way that linguistic representations are not. Thus, the content of a neural representation of the cat being on the mat may be *very* roughly approximated as follows: *cat on mat there now and will likely evolve in such and such a way*. Notice that I didn't use a that-clause, because typical neural representations are not propositional representations but simulations of their target.

A specific content may be distributed over a relatively large ensemble of neurons. Yet content is relatively localized in the sense that it is carried by a specific vehicle born by a specific bearer (neuron/ensemble/circuit) and not diffused through the whole neurocognitive system, or even a large part thereof. Yet content (qua causal power) also depends on the causal role that the firing of a neuron/neural ensemble/neural circuit plays within the neurocognitive system, so it depends on the structural and functional relations between the vehicle (and therefore the vehicle bearer, the neuronal structure) and other relevant portions of the system. Since content is acquired by the neurocognitive system through learning via feedback from the environment (more on this below), it is acquired holistically thanks to the action of a system larger than the bearer of the content, and it depends on the holistic relations between its bearer and the rest of the system for its existence qua content. Yet content is also somewhat localized in the sense of being possessed by a small part of the system in virtue of the specific causal role that subsystem plays within the whole system.

In other words, the content of a neural representation is a manifestation of a causal power (the power to track a target), yet this content is created by a broader learning process involving a larger system, and the fact that it functions as content is made possible by the broader causal role that the content plays in guiding behavior within the system.

In summary, there are three causal processes pertaining to content: the learning process that creates the content, the causal process that defines the content (as tracking a certain target and predicting its evolution), and the causal process that makes it possible for the content to guide behavior.

## THE SITUATEDNESS OF NEURAL REPRESENTATION

For neural representations to exist at all, the system that constructs and maintains them—the neurocognitive system—must be embodied, embedded, enactive, and affective. This situatedness of neural representations is needed because neural representations and the computations that are interdependent with them emerge diachronically through the dynamical interaction between the nervous system, its body, and its environment in a way that must take into consideration the organism's needs. Let's unpack this point, one step at a time.

Neurocognitive systems are made out of neurons and other cells; the neurons, connected into networks, are the main components performing cognitive functions. The structure and functions of neurocognitive systems are innately constrained. The structure and functions of an organism's body affect how its neurocognitive system develops and what processes it performs (Chiel and Beer, 1997). In addition, developmental processes that are at least partially under genetic control determine the differentiation of the neurocognitive system into different systems (cortex, cerebellum, hippocampus, etc.), the formation of different subsystems (cortical areas, columns, nuclei), much of the wiring between systems and subsystems, the main biophysical properties of different types of neurons, the transduction of external stimuli into firing rates within sensory systems, the transduction of firing rates into muscle contractions at the neuromuscular junction, and so forth. All these factors constrain the type of representations and computations neurocognitive systems can perform and the kinds of behaviors they can exhibit (e.g., Kim et al., 2017; Wang et al., 2018). That said, one of the most important features of neurocognitive systems, which is also built through development, is their *plasticity*, that is, their ability to change their structure and functions in response to their dynamic interaction with body and environment. Plasticity is the basis for the ability to *learn*, which in turn allows neurocognitive systems to construct and shape their representations and computations.

The study of how biological neural networks learn has influenced and has been influenced by the study of *artificial* neural networks. Comparing the types of learning that occurs in biological vs. artificial neural networks will help us highlight how important situatedness is to learning in biological neural networks and what might still be missing from current AI technology.

Artificial neural networks can learn in three main ways: supervised, unsupervised, and by reinforcement. *Supervised* learning occurs when an agent external to the network calculates the error produced by the network, uses such error to adjust the structure (and therefore the functions) of the network to improve performance, and repeats this process until the network exhibits the desired performance. This is often done by feeding

[10]Shea (2018, p. 36, 39) has independently argued that content arises out of a combination of a vehicle's intrinsic and relational properties.

the network labeled data during the training period, that is, inputs that already include information about how the network is supposed to classify the data. In contrast, *unsupervised* learning occurs when the network itself adjusts its structure (and therefore its functions) in response to its inputs in order to find, extract, and represent similarities, invariants, and associations within its inputs, without receiving external feedback on how to improve its performance. Finally, *reinforcement* learning occurs when a network performs actions in response to its input, receives a reward signal in response to successful actions, and uses the reward signal to adjust which actions it will select in the future (Sutton and Barto, 2020).

Both supervised and traditional unsupervised learning have limits. Supervised learning is limited by the requirement of labeled data, which may or may not be always available in large enough quantity. Unsupervised learning is limited by the absence of any external information on how the inputs should be processed; thus, it works best for tasks that require merely extracting patterns from inputs. To overcome these limitations, a more recent approach involves a type of unsupervised learning called *self-supervised* learning: artificial neural networks that learn by extracting supervisory signals from the data themselves, without relying on explicit labels supplied by external agents. By relying on the structure of the data, self-supervised learning networks attempt to predict one portion of an input from another portion, and then use any resultant discrepancy to improve their representations and computations. Adding the ability to learn from rewards and punishments turns a neural network into a *reinforcement* learning network, which allows it to learn by trial-and-error how to respond to different situations.

None of the training methods for artificial neural networks are a perfect fit for the type of learning that occurs within neurocognitive systems. Unlike in supervised learning, there are no external agents labeling the data that enter neurocognitive systems or calculating how the structure of neurocognitive systems should be adjusted to improve performance. Therefore, neurocognitive systems do not undergo supervised learning as it occurs in artificial neural networks. In addition, unlike traditional unsupervised learning, neurocognitive systems are not limited to processing their inputs in the absence of external feedback.

The types of AI learning that are closest to what neurocognitive systems do are self-supervised learning and reinforcement learning. Like artificial systems undergoing self-supervised learning, neurocognitive systems can extract structure from their inputs, attempt to predict how the inputs will evolve, and use any discrepancy to improve their representations and computations (cf. (Buckner, forthcoming)). But even self-supervised learning falls short because, in general, self-supervised learning does not involve direct feedback from either the system, the body, or the environment about the effects of the system's actions—if nothing else, because typical artificial neural networks do not act in the world through a body in real time. In contrast, neurocognitive systems are constantly directing their body to act within their environment, use efference copies of their own motor commands to adjust their expectations about how their sensory inputs will change, and collect information about the effects of their motor commands on both body and environment shortly after issuing the commands. Thus, neurocognitive systems can and do use constant, real-time feedback to correct their structure so as to improve their performance.

This lacuna is addressed in part by reinforcement learning. Like artificial systems undergoing reinforcement learning, neurocognitive systems can adjust their action selection by responding to rewards and punishments. There are at least four important differences. First, neurocognitive systems learn in the real world within a relatively short amount of time, whereas current AI techniques are too inefficient to learn realistic tasks in the real world within a reasonable time; learning occurs within simulated worlds and then the acquired knowledge may be transferred to the real world with some degree of success (OpenAI et al., 2019a,b). Second, neurocognitive systems—unlike ordinary artificial neural networks—include an *internal* system of evaluative signals, so neurocognitive systems are not limited to learning from external evaluative signals. Third, neurocognitive systems use several different types of internal reward and punishment signals instead of just one type of evaluative signal. Fourth, insofar as neurocognitive systems can learn from external evaluative signals, such as a parent or teacher telling them "Yes" or "No," they have to first learn to interpret such signals. To distinguish the type of learning that neurocognitive systems engage in from standard AI techniques, I will call it *active* learning[11].

The most important feature that active learning shares with AI methods is that *the learning process itself shapes the computations at the same time that it builds the representations*. This marks a critical difference from conventional computers. In conventional computers, the processor manipulates data in accordance with instructions, its circuitry usually remains the same over time, while instructions and data are stored in separate memory registers. Computer instructions have internal semantic content that correspond to the operations performed by the processor, while data can mean anything at all—their content need not have anything to do with the computational operations performed on them. Usually, the operations performed on data match their contents, but this happens only because programmers and users ensure that they do. In fact, computer data need not even mean anything at all. Because of this, if computer data have semantic content at all, as they usually do, they have *derivative* content.

In contrast, in learning neural networks, the operations performed by the units are what activates their representational states, and the representational properties of the states are what allows the network to perform subsequent computational operations efficiently. This mutual dependence exists because both the computational operations and the representations are constructed, jointly and at the same time, by one and the same learning process (cf. Shea, 2018, p. 217). As a result, *within learning neural networks, computations and representations are mutually constitutive of each other and, thus, automatically*

[11]The label "active learning" is used in pedagogy for a method of learning in which students are not merely listening to lectures or reading material (i.e., passive learning) but are actively engaged with the material through discussions, writing assignments, role play, etc. I am repurposing this label for the type of learning that neurocognitive systems spontaneously engage in.

*coordinated*. This is not enough to conclude that neural networks have original, causally efficacious semantic content, but we will soon see that it is an important step in that direction[12]. What is also needed in order to acquire original semantic content is that the network be embodied, embedded, and enactive.

Before getting there, I want to point out another, underappreciated difference between neurocomputational systems and conventional computers. Within conventional computers, the only kind of information processing that takes place is the computation of outputs based on inputs and internal states. In contrast, neurocomputational systems are constantly engaged in two types of information processing at once. Like conventional computers, they yield outputs as a function of their inputs and internal states. *Unlike* conventional computers, they also learn—that is, they use a number of information sources together with their self-organizing capacity to alter their structure and, therefore, their future functions.

It's worth pointing out what sorts of information sources neurocomputational systems can use to actively learn. They include the timing and frequency of their vehicles (primarily, neuronal spikes), the channels through which input signals arrive (visual, auditory, olfactory, etc.), the correlation between one portion of a signal and another portion, and the dependencies between various sorts of input signals (from the environment, body, or neurocognitive system itself, as in the case of efference copy), internal states (such as internal states of the simulations of body and environment and the internal evaluative signals they elicit), and output signals (such as motor commands). By exploiting these relationships and performing operations that are sensitive to them, a neurocognitive system can process information using medium-independent vehicles. In addition, by exploiting the different patterns of dependencies that occur between internal signals and signals from the body, on one hand, and between internal signals and signals from the environment, on the other hand, neurocognitive systems can learn to distinguish between their body and their environment. The upshot is that neurocognitive systems can build representations with original semantic content because neural representations and the computations that manipulate them are functions not only of each single network's inputs and internal states but also of the real-time dependencies between different portions of the whole neurocognitive system's inputs as well as between inputs, internal states, and outputs, which in turn carries information about the body and environment of the system.

Thus, active learning requires *embodiment*—that is, a tight dynamic coupling between neurocognitive system and body[13]. This is true not only because the body contains the sensors and effectors that neurocognitive systems need in order to receive information and act on it. It's also because the real-time feedback loop between neurocognitive systems and their body, whereby the body moves in direct response to neural activity and almost immediately sends information back to the neurocognitive system about how it's moved, is needed for the neurocognitive system to learn how to represent its body, how to represent its body *distinctly* from its environment, and how to effectively simulate and control its body. Since the body is, in turn, the main receiver of information about the environment, the neurocognitive system could not fulfill its learning potential—much less learn how to direct its body within its environment by using internal simulations as a guide—without its constant dynamic interaction with its body.

Active learning requires *embeddedness* as well—that is, a tight dynamic coupling between nervous system, body, and environment. This is true not only because the environment contains the sources of information most senses are sensitive to (except for proprioception, which is perception of the body itself) or because the body itself could not function in the absence of its environment. It's also because the real-time feedback loop between neurocognitive systems and their environment—whereby the environment mostly remains the same regardless of the organism's movements even while the perspective of the organism changes, and yet the environment also changes in specific ways that depend on the actions performed by the organism—is needed for the neurocognitive system to learn how to represent its environment, how to represent its environment *distinctly* from its body, and how to effectively simulate and act within its environment. For example, abnormal visual stimulation during a developmentally critical period impairs vision in ways that can be irreversible (e.g., Hubel and Wiesel, 1970). The neurocognitive system cannot develop properly and cannot fulfill its learning potential without dynamically interacting with its environment, in a way that is mediated by its body.

Active learning requires *enaction* too. For present purposes, enaction is a kind of dynamic interdependence of a system and its environment that unfolds continuously in real time. Specifically, when enaction occurs, cognitive states and processes affect the organism's body and environment while the body and environment affect cognitive states and processes (cf. what Ward et al., 2017 call "sensorimotor enactivism"). Enaction in this sense is already largely implicit in what I said above—let's highlight its most relevant aspects. At any given time, the neurocognitive system is building and updating a simulation of its body and environment and using such a simulation to guide behavior. Meanwhile, each motor command affects (i) how the body moves, (ii) how the sensory input changes (if nothing else, because the position of the body relative to its environment changes), and (iii) some ways that the environment changes (because the organism's actions change it). Moreover, the simulation is attempting to predict how all of this is about to unfold, and the system compares its predictions to its sensory data. Sensory data, in turn, are the main way that the environment affects neurocognitive systems in real time. All of

[12]For an independently developed yet converging argument that the coordination between vehicles and their content helps solve the problem of the causal efficacy of content see Shea (unpublished).

[13]Different authors characterize embodiment, embeddedness, and enaction in different ways (for a recent review, see Shapiro and Spaulding, 2021). Some authors define embodiment, embeddedness, or enaction in ways that preclude computation and representation (e.g., Thompson, 2007, p. 13); this begs the question at hand. I adopt characterizations that are present in the literature, do not beg the question of computation and representation, and suit present purposes. A fuller treatment of the relation between the present argument and the many themes from the literature on situated cognition will have to wait for another occasion.

the dependencies between sensory inputs and motor actions are constantly exploited by neurocognitive systems to update their internal simulations as well as to learn how to improve their simulations and the way they guide behavior.

Finally, active learning requires *affect*. In the most basic sense, affect is a system of internal signals that evaluate the state of the organism and its environment to motivate the selection of actions that satisfy the organism's needs. Animals, or at least animals of sufficient complexity and behavioral flexibility, need affect in this sense to select actions that satisfy their needs as well as evaluate external situations and, eventually, learn to select action sequences that are adaptive within different situations. As we have seen, affect in this sense is an aspect of reinforcement learning, which is an aspect of active learning.

As a result of the dependence of active learning on the situatedness of neurocognitive systems, neural representations and computations themselves are embodied, embedded, enactive, and affective. That is, neural representations and computations are the result of the tight interdependence between neurocognitive systems and their body and environment—neurocognitive systems track their targets and guide behavior thanks to their situatedness.

This account is a kind of content externalism, to the effect that neural representations require a direct dynamical coupling to the body and environment in order to exist at all as well as to acquire their semantic content. Content is determined in part by the environment together with the interaction between the nervous system and its environment. As a consequence, neural representations are individuated at least in part by the external variables they have the function to track. This accords with standard definitions of content externalism (Rowlands et al., 2020).

This content externalism is a close relative of but should not be confused with the traditional content externalism defended by Putnam 1975. According to traditional content externalism (adapted to neurocognitive systems), a difference between two environments that is undetectable by the organism, such as a difference in chemical composition between two substances that the organism has no sensory ability to discriminate, is enough to alter the semantic content of a representation. For example, suppose that an organism *A* has learned to activate representations of type *R* in the presence of substance *S* so as to guide behavior with respect to *S*. In light of teleosemantics, tokens of *R* represent *S*. Suppose that organism *A* has an exactly similar duplicate A* who lives in an environment where substance *S** is present in exactly the same contexts in which *S* is present within *A*'s environment, yet neither the original organism *A* nor its duplicate *A** has any way to distinguish *S** from *S*. As a result, within the duplicate *A**, tokens of *R* get activated in the presence of *S**. Traditional content externalism maintains that, in the duplicate A*, tokens of *R* represent *S** rather than *S*.

Traditional content externalism is neither needed nor plausible within the kind of naturalistic perspective I advocate. The sort of case envisioned by traditional content externalism is an exotic case that is unlikely to occur in real life. If it were to occur, the reasonable thing to say is that there are two types of substances, *S* and *S**, represented by tokens of *R*. A real-world example is the gemstone jade, which may be composed of either of two chemically different minerals, jadeite and nephrite[14]. Prior to modern chemistry, no one knew that there were two types of jade. Nevertheless, then as now, and contrary to traditional content externalism, the term "jade" does not mean just jadeite or just nephrite depending on whether we are looking at jadeite or nephrite, or whether we are in an environment where only jadeite is present or only nephrite is present, or, as traditional content externalists would put it, whether we live on a planet where just jadeite or just nephrite is present. "Jade" just means jade, i.e., something that can be either jadeite or nephrite. By the same token, neural representations represent what neurocognitive systems use them to track, regardless of how many different types of underlying structures activate the same representation[15].

In conclusion, neural representations emerge diachronically through the dynamical interaction between neurocognitive systems, their body, and their environment, and they depend on such a dynamic interaction for their existence and updating. This situatedness of neural representations allows us to solve the problems of content.

## HOW SITUATEDNESS SOLVES THE PROBLEMS OF CONTENT

The first problem is the source of original semantic content: how do neural representations acquire original (i.e., non-derivative) semantic content? The situatedness of neurocognitive systems is the very source of their representations' original content. As we've seen, original content itself emerges via a combination of biological evolution and active learning from the constant interaction between nervous system, body, and environment. The original content of a neural representation is a property acquired by the representation via a combination of evolution shaping development and active learning that the system undergoes as it constructs internal simulations of its body and environment to guide the organism's behavior. Perceptual representations and their original content may be more dependent on receiving sensory information than on guiding action, while the reverse may be true of motor representations; nevertheless, for all types of neural representations to be normally acquired and coordinated, all the forms of situatedness we discussed must contribute[16].

The second problem is the coordination between vehicles and their content: how do vehicles and contents get matched with one another so that the computational operations the nervous system performs over the vehicles match their semantic content? Situatedness solves the coordination problem because the contents themselves are an aspect of the vehicles' functional role, and such a functional role (including the computational

[14]Jade was discussed extensively in the debate on reductionism about mental states (Kim, 1992; Fodor, 1997). I am putting the example to a different use.

[15]This conclusion is consistent with many critiques of traditional content externalism; see Sections 3.2 and 3.3 of Rowlands et al. 2020 for a review.

[16]Thanks to a referee for pointing out that there may be cases of atypical development, neurodiversity, or neuropathology in which some forms of situatedness do not contribute to the development of neural representations.

operations to be performed on the vehicles) are learned by the system via its interaction with body and environment and the feedback it receives through that interaction at the same time that the content itself is acquired. As we've seen, in neurocognitive systems there is no separation between the semantic content of a neural representation and the computational operations performed over them. The computational operations are an aspect of what gives a neural representation its content; neural representations with the content they have are what allows subsequent computational operations to be performed; the contents and the computational operations are acquired together as the system undergoes active learning.

Let's consider this a bit further. When a neurocognitive system begins to develop, it possesses some ability to process its inputs, build internal states, deliver outputs, and learn from the feedback it receives from itself, its body, and its environment. These initial operations may be partially random but they are also constrained by the architecture and biophysical properties of the system, the morphology and organization of the body, and the structure of the environment. The system may already have a system of internal representations built by developmental processes, or it may be closer to a blank slate. If the system does have innate internal representations it must be because evolutionary and developmental processes sufficiently analogous to active learning have constructed them so that their semantic content matches the operations performed by the system over them or else such representations could not function as such. Over time, it is precisely the process of dynamic interaction between nervous system, body, and environment that allows the system to acquire new or more sophisticated representations at the same time that it learns how to use them. Thus, the neural representational vehicles and their content can only arise together because they are two sides of the same coin. The matching between vehicles and contents is guaranteed by the fact that both contents and the operations performed over the vehicles are joint products of the same active learning process.

The third problem is the causal efficacy of content: how can the semantic content of neural representations be causally efficacious? Situatedness solves the causal efficacy problem because, as we've seen, the content of situated neural representations is an aspect of the causal powers of its vehicles. As a result, unlike typical artificial computing systems, neurocomputational systems are sensitive to the semantic content of their vehicles.

To illustrate, consider a token *r* of neural representation type *R*. Suppose that the system has actively learned to activate tokens of *R* to track and simulate the presence of dogs in its environment and guide behavior with respect to dogs (e.g., Bracci et al., 2019). According to the version of informational teleosemantics I advocate, *r* has original semantic content that can be expressed, approximately, by *dog there now and will likely evolve in such and such a way*. Such a content is not something distinct from and independent of *r*'s causal powers. Rather, *r*'s content is both a manifestation of some of the neurocognitive system's causal powers and a causal power of its own, which can trigger further manifestations. In this case, *r*'s content is an aspect of its power (again, within the context of the neurocognitive system) to guide the system's behavior with respect to a dog being there now. This is made possible by the automatic coordination of *r* (and the computations that process *r*) with *r*'s content that is created when the disposition to activate tokens of *R* is constructed within the system via active learning. Thus, *r*'s content causes the system's behavior with regards to a dog being there now. This is how the semantic content of neural representations causally explains behavior.

The fourth problem is the indeterminacy of content: how can neural representations be said to have semantic content when theorists can't agree on what content they have? Situatedness solves the problem of the indeterminacy of content because the content of the kind of basic neural representations we've been discussing need not have fully determinate semantic content like declarative sentences within a human language. Neural representations have the kind of content that the system needs in order to guide behavior; the kind of content that is ecologically significant and that evolution can act on. The type of behavior depends on the type of organism, and the content of individual neural representations is for neuroscientists to investigate empirically, not for philosophers to intuit about. By investigating the response properties of neurons and neuronal populations, neuroscientists can determine what such neurons or populations are most responsive to under relatively good sensory conditions, and that is their semantic content. If there are different, nonequivalent ways of labeling such contents linguistically (e.g., "fly," "small dark moving entity," "food"; or "*S*" versus "*S**"), this doesn't matter so long as all such labels are extensionally equivalent within the relevant ecological niche[17]. Only when it comes to linguistic cognition do the very special neurolinguistic systems that are involved acquire the kind of categorical contents that admit of full-blown truth conditions. How to get there is a complex story that still needs to be told in detail (some hints are provided in Piccinini and Hetherington, unpublished; Piccinini, 2020b, 2022).

The fifth problem is the distality of content: why should the distal stimulus be the content of a neural representation rather than any of its more proximal stimuli? Situatedness solves the distality problem because different items along the causal chain from distal stimulus to neural representations exhibit different patterns of dependency. As we've seen, neural representations are not static—they dynamically predict the evolution of their target and guide behavior with regards to the target. Meanwhile, the system obtains and processes feedback in response to its actions. At the very least, the organism's movements, including its eye movements, constantly change the precise point of view from which the nervous system obtains sensory data from any given target. The dependency patterns between different items along the causal chain from distal stimulus to internal states are different, the different items evolve in different ways, and changing point of view alters them in different ways. Therefore, as

[17]Whether the labels in our frog example are actually extensionally equivalent within the ecological niche of frogs is questionable. For instance, frogs eat way more than just flies, so "fly" is not extensionally equivalent to "food". At any rate, we should let neuroscientists find the best way to characterize the content of neural representations.

soon as the system attempts to predict how something will evolve over time and improve its predictive performance as well as its action efficacy, it must extract the invariants that are relevant to external stimuli—those it might have to interact with—and discard any invariants that pertain to more proximal stimuli as spurious. It is part and parcel of a neural systems' active learning to acquire representations capable of predicting the evolution of the distal stimuli—those the system might actually interact with[18].

The sixth and last problem is the possibility of misrepresentation: how can a neural representation misrepresent a stimulus that triggers it? Traditional information teleosemantics is often thought to provide a satisfactory account of misrepresentation. The solution is supposed to be that misrepresentation is failure to perform representational function. This is a huge step in the right direction but it's not enough by itself. The problem with this standard solution is that it requires determining representational function precisely enough to make room for misrepresentation. Specifically, there must be something that determines what each internal state has the function to represent so that, when the state responds to something else, misrepresentation ensues. Extant proposals are that either evolution (Neander, 2017) or an appropriate learning period (Dretske, 1988) determine what each state has the function to represent. I already ruled out evolution as the source of the right notion of function, so that's a nonstarter. As to learning, no one has found a principled way to distinguish the learning period from the rest of the life of a representational state, such that after the learning period is over the representational function is fixed. In some cases, there is a critical learning period that may be the basis for establishing the representational functions of internal states. But, in general, neurocognitive systems never stop learning!

Situatedness comes to the rescue because, again, different stimuli engage in different patterns of dependencies. The most obvious difference is feedback in response to the organism's actions. Again, neural representations are dynamical simulations of their environment, which are largely learned. At any given time, neurocognitive systems have multiple representations that could be activated in response to incoming sensory data. Suppose that, during a dark night, in response to a stimulus, a system activates a COW-representation—that is, the kind of representation it has learned to activate when it needs to simulate cows. The COW-representation yields specific predictions about how the sensory data will change if the stimulus is approached (i.e., it will look more and more distinctly like a cow), or if the stimulus makes a vocalization (i.e., it will "moo"), or what their footprints will look like, and so forth. As soon as enough sensory feedback is collected that matches a different representation better than the current one, the system itself should self-correct, and it will self-correct if it's functioning properly. That is, the system will deactivate the COW-representation and activate one that fits the sensory data better—e.g., a HORSE-representation. Thus, misrepresentation occurs when a system activates a representation, targeting a stimulus, which makes worse predictions about incoming data about what a stimulus will do and how it will appear under various possible conditions than an alternate representation that is also available to the system. In short, misrepresentation arises from the interaction of learning, simulation, and the ability to detect errors and make corrections. The ability of neurocognitive systems to correct their own misrepresentations is also another way of seeing that their content is causally efficacious (Bielecka and Miłkowski, 2020)[19].

## CONCLUSION

I have argued that, far from being opposites as so many have thought, situatedness and representation are more deeply intertwined than anyone suspected. What makes neural representations possible is the very situatedness of the processes that acquire neural computations and representations.

Neurocognitive systems are indeed embodied, embedded, affective, dynamically interact with their environment, and use feedback from their interaction to acquire their own representations and computations via active learning. This accounts for the following: (1) neurocognitive systems construct neural representations with original semantic content, (2) their neural vehicles and the way they are processed are automatically coordinated with their content, (3) such content is a special kind of causal power and hence causally efficacious, (4) is determinate enough for the system's purposes, (5) represents the distal stimulus, and (6) can misrepresent. This proposal hints at what artifacts should be like in order to acquire the basic cognitive abilities possessed by neurocognitive systems.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## ACKNOWLEDGMENTS

This work was partially done on the territories of the Kickapoo, Kaskaskia, Myaamia, Ogaxpa, and Osage peoples. Thanks to Adam Safron for inviting me to submit to this special issue; this paper would probably not exist without that invite. In researching this paper, I am most indebted to two people. First, Michael Carl, director of the Center for Research and Innovation in Translation and Translation Technology at Kent State University, kindly invited me to give a talk in the hope that I would help clear up some of the confusion surrounding computationalism and 4EA approaches; that talk was a major impetus for writing this

[18]Some recent teleosemantic literature moves at least part of the way in the same direction (e.g., Neander, 2017, Ch. 9; Garson, 2019; Schulte, 2021).

[19]On the role of error correction, I'm also indebted to Gładziejewski (2015).

paper and I thank the audience at Kent State for their feedback. Second, Eric Thomson taught me more than anyone else about neural representation and gave me invaluable comments on a previous version of this paper that led to many improvements. Thanks also to Giuseppe Artese, Oliver Casper, Joe Dewhurst, Kay Fitzpatrick, Paweł Gładziejewski, Nick Shea, Zina Ward, and especially Jonny Lee and Marcin Miłkowski for helpful comments on previous versions of this paper. Thanks to Matej Kohar for an illuminating conversation about mechanisms and representations.

## REFERENCES

Adams, F., and Aizawa, K. (2021). "Causal theories of mental content," in *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), eds. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/fall2021/entries/content-causal/

Anderson, N. and Piccinini, G. (unpublished). *The Physical Signature of Computation: A Robust Mapping Account of Implementation*.

Bielecka, K., and Miłkowski, M. (2020). "Error detection and representational mechanisms," in *What are Mental Representations?,* eds. J. Smortchkova, K. Dołega, and T. Schlicht (Oxford: Oxford University Press), p. 287–313

Bracci, S., J. B., Ritchie, I., Kalfas, H., and Op de Beeck, P. (2019). The ventral pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *J. Neurosci.* 39, 6513–6525. doi: 10.1523/JNEUROSCI.1714-18.2019

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese* 195, 5339–5372. doi: 10.1007/s11229-018-01949-1

Buckner, C. (forthcoming). A Forward-Looking Theory of Content. *Ergo*.

Casper, M.-O., and Artese, G. F. (2020). Maintaining coherence in the situated cognition debate: what computationalism cannot offer to a future post-cognitivist science. *Adapt. Behav.* 2021: 1059712320967053. doi: 10.1177/1059712320967053

Chiel, H. J., and Beer, R. D. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci.* 20, 553–557. doi: 10.1016/S0166-2236(97)01149-1

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: MIT Press.

Dewhurst, J., and Villalobos, M. (2018). Enactive autonomy in computational systems. *Synthese* 195, 1891–1908. doi: 10.1007/s11229-017-1386-z

Dretske, F. (1986). "Misrepresentation," in *Belief: Form, Content, and Function*, ed. R. Bogdan (Oxford: Clarendon), 17–36.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Facchin, M. (2021). Structural representations do not meet the job description challenge. *Synthese* 199, 5479–5508. doi: 10.1007/s11229-021-03032-8

Fodor, J A. (1994). *The Elm and the Expert: Mentalese and Its Semantics.* Cambridge, MA: MIT Press.

Fodor, J. A. (1997). Special sciences: still autonomous after all these years. *Philos. Perspect.* 11, 149–163. doi: 10.1111/0029-4624.31.s11.7

Garson, J. (2019). Do constancy mechanisms save distal content? *Philos. Q.* 69, 409–417. doi: 10.1093/pq/pqy032

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: a mechanistic perspective. *Stud. Logic Gramm. Rhetoric* 40, 63–90 doi: 10.1515/slgr-2015-0004

Gładziejewski, P., and Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biol. Philos.* 32, 337–355. doi: 10.1007/s10539-017-9562-6

Haugeland, J. (1998). "The intentionality all-stars," in *Having Thought*, eds J. Haugeland (Cambridge, MA: Harvard University Press), 127–170.

Hubel, D. H., and Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J. Physiol.* 206, 419–436. doi: 10.1113/jphysiol.1970.sp009022

Jacob, P. (2019). "Intentionality," in *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), ed. E. N. Zalta. Available online at: https://plato.stanford.edu/archives/win2019/entries/intentionality/

Kauffman, S. (1993). *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford: Oxford University Press.

Kauffman, S. (2002). *Investigations*. Oxford: Oxford University Press.

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philos. Phenomenol. Res.* 52, 1–26. doi: 10.2307/2107741

Kim, S. M., Su, C.-Y., and Wang, J. W. (2017). Neuromodulation of innate behaviors in *Drosophila*. *Annu. Rev. Neurosci.* 40, 327–348. doi: 10.1146/annurev-neuro-072116-031558

Lee, J. (2018). Structural representation and the two problems of content. *Mind Lang.* 34, 606–626. doi: 10.1111/mila.12224

Lee, J. (2021). Rise of the swamp creatures: reflections on a mechanistic approach to content. *Philos. Psychol.* 34, 805–828. doi: 10.1080/09515089.2021.1918658

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proc. IRE.* 47, 1940–1951. doi: 10.1109/JRPROC.1959.287207

Martin, C. B. (2008). *The Mind in Nature*. Oxford: Oxford University Press.

Miłkowski, M. (2017). Situatedness and embodiment of computational systems. *Entropy* 19, 162. doi: 10.3390/e19040162

Miłkowski, M. (2021). Correspondence theory of semantic information. *Br. J. Philos. Sci.* doi: 10.1086/714804

Millikan, R. G. (2021). Neuroscience and teleosemantics. *Synthese.* 199, 2457–2465. doi: 10.1007/s11229-020-02893-9

Montévil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *J. Theor. Biol.* 372, 179–191. doi: 10.1016/j.jtbi.2015.02.029

Morgan, A., and Piccinini, G. (2018). Towards a cognitive neuroscience of intentionality. *Minds Mach.* 28, 119–139. doi: 10.1007/s11023-017-9437-2

Neander, K. (2017). *A Mark of the Mental: In Defense of Informational Teleosemantics.* Cambridge, MA: MIT Press.

Neander, K., and Schulte, P. (2021). "Teleological Theories of Mental Content," in *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), ed. E. N. Zalta. Available online at: https://plato.stanford.edu/archives/spr2021/entries/content-teleological/

OpenAI, Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., et al. (2019b). *Solving Rubik's Cube with a Robot Hand. arXiv*:1910.07113 v0711

OpenAI., Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., and McGrew, B. (2019a). Learning dexterous in-hand manipulation. *Int. J. Robot. Res.* 39, 3–20. doi: 10.1177/0278364919887447

Piccinini, G. (2004). Functionalism, computationalism, and mental contents. *Can. J. Philos.* 34, 375–410. doi: 10.1080/00455091.2004.10716572

Piccinini, G. (2020a). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford: Oxford University Press.

Piccinini, G. (2020b). "Nonnatural Mental Representation," in *What Are Mental Representations?,* eds. K. Dolega, T. Schlicht, J. Smortchkova (Oxford: Oxford University Press), 254–286. doi: 10.1093/oso/9780190686673.003.0010

Piccinini, G. and Hetherington, S. (unpublished). Knowing That as Knowing How: A Neurocognitive Account.

Piccinini, G. (2022). *An Egalitarian Account of Composition and Realization*. The Monist. 105, 276–292.

Poldrack, R.A. (2021). The physics of representation. *Synthese.* 199, 1307–1325. doi: 10.1007/s11229-020-02793-y

Putnam, H. (1975). *The Meaning of 'Meaning.' Philosophical Papers, Vol. II: Mind, Language, and Reality.* Cambridge: Cambridge University Press.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge, UK: Cambridge University Press.

Ramsey, W. M. (2016). Untangling two questions about mental representation. *New Ideas Psychol.* 40, 3–12. doi: 10.1016/j.newideapsych.2015.01.004

Rowlands, M., Lau, J., and Deutsch, M. (2020). "Externalism about the mind," in *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), ed. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/win2020/entries/content-externalism/

Schulte, P. (2021). Constancy mechanisms and distal content: a reply to Garson. *Philos. Q.* 72, 229–223. doi: 10.1093/pq/pqab008

Shapiro, L., and Spaulding, S. (2021). "Embodied Cognition," in *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), ed. Edward N. Zalta. Available online at: https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/

Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Shea, N. (unpublished). Content-Specific Transitions Between Concepts Answer the Overdetermination Challenge.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Sutton, R. S., and Barto, A. G. (2020). *Reinforcement Learning: An Introduction, Second Edition*. Cambridge, MA: MIT Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Thomson, E. E., and Piccinini, G. (2018). Neural representation observed. *Minds and Machines* 28, 1–45. doi: 10.1007/s11023-018-9459-4

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems* 5, 187–196. doi: 10.1016/0303-2647(74)90031-8

Wang, Y., Wang, L., Xu, Q., Liu, D., Chen, L., Troje, N. F., He, S., and Jiang, Y. (2018). Heritable aspects of biological motion perception and covariation with autistic traits. *Proc. Natl. Acad. Sci.* 115, 1937–1942. doi: 10.1073/pnas.1714655115

Ward, D., Silverman, D., and Villalobos, M. (2017). Introduction: the varieties of enactivism. *Topoi* 36, 365–375. doi: 10.1007/s11245-017-9484-6

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Piccinini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 14 April 2022
doi: 10.3389/fnbot.2022.840658

# Embodied Object Representation Learning and Recognition

*Toon Van de Maele*, Tim Verbelen, Ozan Çatal and Bart Dhoedt*

*IDLab, Department of Information Technology, Ghent University - imec, Ghent, Belgium*

Scene understanding and decomposition is a crucial challenge for intelligent systems, whether it is for object manipulation, navigation, or any other task. Although current machine and deep learning approaches for object detection and classification obtain high accuracy, they typically do not leverage interaction with the world and are limited to a set of objects seen during training. Humans on the other hand learn to recognize and classify different objects by actively engaging with them on first encounter. Moreover, recent theories in neuroscience suggest that cortical columns in the neocortex play an important role in this process, by building predictive models about objects in their reference frame. In this article, we present an enactive embodied agent that implements such a generative model for object interaction. For each object category, our system instantiates a deep neural network, called Cortical Column Network (CCN), that represents the object in its own reference frame by learning a generative model that predicts the expected transform in pixel space, given an action. The model parameters are optimized through the active inference paradigm, i.e., the minimization of variational free energy. When provided with a visual observation, an ensemble of CCNs each vote on their belief of observing that specific object category, yielding a potential object classification. In case the likelihood on the selected category is too low, the object is detected as an unknown category, and the agent has the ability to instantiate a novel CCN for this category. We validate our system in an simulated environment, where it needs to learn to discern multiple objects from the YCB dataset. We show that classification accuracy improves as an embodied agent can gather more evidence, and that it is able to learn about novel, previously unseen objects. Finally, we show that an agent driven through active inference can choose their actions to reach a preferred observation.

**Keywords: generative modeling, robotic perception, deep learning, active inference, representation learning**

OPEN ACCESS

***Edited by:***
*Andy Clark,*
*University of Sussex, United Kingdom*

***Reviewed by:***
*Subutai Ahmad,*
*Numenta Inc., United States*
*Karl Friston,*
*University College London,*
*United Kingdom*

****Correspondence:***
*Toon Van de Maele*
*toon.vandemaele@ugent.be*

***Received:*** *21 December 2021*
***Accepted:*** *23 February 2022*
***Published:*** *14 April 2022*

***Citation:***
*Van de Maele T, Verbelen T, Çatal O and Dhoedt B (2022) Embodied Object Representation Learning and Recognition. Front. Neurorobot. 16:840658. doi: 10.3389/fnbot.2022.840658*

## 1. INTRODUCTION

Having a machine understand the world from pixels has been a long standing challenge defining the field of computer vision (Hanson, 1978). In the last decade, we have witnessed a proliferation of deep learning techniques in this domain, which started with the leap in performance obtained by a convolutional neural network (CNN) on object classification (Krizhevsky et al., 2012). Besides the exponential scaling of available compute resources, this progress is mainly fueled by the collection of massive datasets like ImageNet (Deng et al., 2009). The main strength of these techniques is that their classification accuracy typically improves as they are trained on more data, scaling to datasets containing billions of images (Mahajan et al., 2018). However, this strength is also becoming a

main point of critique, as an exponential increase in compute (and energy) resources is required for marginal gains (Thompson et al., 2021). Moreover, these classifiers are known to be vulnerable to ambiguous and adversarial samples (Gilmer et al., 2018), and are restricted to object categories known and seen during training.

Humans on the other hand are embodied agents (Safron, 2021), allowing them to resolve ambiguity by actively sampling the world (Mirza et al., 2018). They are also much better learners: by the age of two, toddlers can recognize around 300 object categories (Frank et al., 2016), and can generalize a newly learned label to instances they have never seen before (Landau et al., 1988). Moreover, toddlers actively engage with their environment, visually exploring objects from various viewpoints by looking at and playing with them (James et al., 2014). In contrast to datasets collected for machine learning, which aim to collect a large and diverse set of exemplars of each object category, toddlers rather learn from a severely skewed data distribution, where only a small set of object instances are pervasively present, yet still we are able to generalize (Clerkin et al., 2017). Therefore, we propose a more enactive method for object category learning, in which an artificial agent can actively sample viewpoints.

Predictive coding is a paradigm based on the hypothesis of the Bayesian brain (Rao and Ballard, 1999), which makes the assumption that cortical circuits perform Bayesian inference to find the hidden causes of the observed signals. According to this paradigm, the brain entails a generative model and uses this to encode the error on the predicted observation.

Active inference is a process theory of sentience, which states that intelligent systems build a generative model of their world and act by minimizing a bound on surprise, i.e., the variational free energy (Friston et al., 2016). As such, active inference can not only be used to build artificial agents (Çatal et al., 2020a), but also to develop theories about functioning of the brain (Parr and Friston, 2018). For instance, Parr et al. (2021) propose an active inference account for human vision, which considers perception as inferring a scene as a factorization of separate (parts of) objects, their identity, scale and pose. Factorizing object identify from their scale and pose is consistent with the so called two stream hypothesis, which states that visual information is processed by a dorsal ("where") stream on the one hand, representing where an object is in the space, and a ventral ("what") stream on the other hand, representing object identity (Mishkin et al., 1983).

Similarly, Hawkins et al. (2017) hypothesize that cortical columns in the neocortex build object-centric models, capturing their pose in a local reference frame, encoded by cortical grid cells. Also empirical evidence from cognitive psychology showed that humans, given a single view of an object never seen before, have strong expectations about rotated views of that object, implying internal representations of three dimensional objects rather than two dimensional views (Tse, 1999). Recent findings in recordings of rhesus monkey brains provide evidence that indeed 3D shape is encoded in the inferior temporal cortex (Janssen et al., 2000).

Drawing inspiration from all these findings, we present a system for learning object-centric representations from pixel data. Akin to how a toddler interacts with a toy, we devise an artificial agent that can look at a 3D object from different viewpoints in a simulated environment. Parallel to cortical columns, our system learns separate models, which we call Cortical Column Networks (CCN) for separate object categories, which encode object pose and identity in two separate factors. An ensemble of CCNs then forms the agent's generative model, which is optimized by minimizing free energy. By engaging in active inference, our agent can realize preferred viewpoints for certain objects, while also resolving ambiguity on object identity.

Building on previous work (Van de Maele et al., 2021a), we now evaluate our agent on pixel data rendered from 33 objects from the YCB benchmarking dataset (Calli et al., 2015). In this article, we show that using object-specific models introduces the ability to classify out-of-distribution objects through a two-stage process that first aggregates the votes and then compares the prediction error on the likelihood of the observation. We devise a mechanism to aggregate information over multiple observations, and show that an embodied, enactive agent outperforms a static classifier for the object classification task. Moreover, we provide qualitative insights on how the system resolves ambiguity through the predictive model.

Additionally, we illustrate how the agent can be drawn to preferred observations through the active inference paradigm, which is crucial for object interactions such as grasping. We investigate the behavior of the latent code representing the object pose and show that the model maps similar observations to the same latent, leveraging symmetrical properties of the object structure to reduce the model complexity.

To summarize, the contributions of this article are threefold:

- We propose an object-centric model (CCN) that learns separate identity and pose factors directly from pixel-based observations through the minimization of free energy. The ensemble of CCNs for known objects form the agents generative model.
- We combine the learned identity latent representation with the likelihood of a CCN to classify objects of both seen (exact identity) and unseen (other class) categories.
- We show that through active inference, the agent can be driven toward an expected observation. We find that the agent reduces complexity in its internal model by mapping similar observations to a similar latent code.

## 2. METHODS

In this section, we first discuss recent generative models for human vision, and propose our generative model for object recognition and perception. Second, we derive the free energy functional to optimize such a generative model under active inference. Finally, we present a particular instance of such a model, using an ensemble of modular deep neural networks, called Cortical Column Networks.

### 2.1. Generative Models for Vision

The Bayesian brain hypothesis finds its origin in the writings of von Helmholtz (1977), and makes the assumption that the

intelligent brain reasons about the world and its uncertainty as a Bayesian process. This perspective is further formalized in terms of active inference, which posits that the brain entertains a generative model of how sensory data are generated, and functions by maximizing a lower bound on Bayesian model evidence through learning and action selection (Friston et al., 2016). Perception then boils down to inverting this model and finding the likely causes that generated the sensory data, i.e., using (approximate) Bayesian inference to compute posterior probabilities over hidden causes.

In the context of vision, this calls for inferring the causes that generate a retinal image in the case of a human, or an array of camera pixels in the case of a machine. Such a generative model should then be able to construct a scene and predict "what would I see if I looked over there" (Mirza et al., 2016). Rao and Ballard (1999) formalize a generative model for vision, through the predictive coding paradigm, by applying the underlying assumption that the external environment generates natural signals in a hierarchical manner by interacting with hidden physical causes such as object shape, texture or luminance. While their generative model considers a factorization in separate latent terms, it does not consider the influence of the observers pose and does not explicitly factorize the scene in separate objects.

A detailed generative model of human vision is proposed by Parr et al. (2021), as schematically represented in **Figure 1**. To predict a retinal image, one needs to know the scene and its constituent objects or entities, as well as the observer's viewpoint within that scene. This is depicted in **Figure 1A**: the observer's viewpoint $\mathbf{v}_t$ at timestep $t$ is determined by its location $\mathbf{l}_t$ and head direction $\mathbf{h}_t$ in the scene $\mathbf{s}$. What the observer sees are the different entities $\mathbf{e}_i$ that are described by their identity $\mathbf{i}$ and their placement in an allocentric reference frame defined by a translation $\mathbf{t}_i$ and rotation $\mathbf{r}_i$. The retinal image $\mathbf{o}_t$ is then formed from the different entities $\mathbf{e}_i$, the observer's viewpoint $\mathbf{v}_t$ together with the context $\mathbf{c}$, e.g. the lighting conditions etc. Importantly, the observer can take action $\mathbf{a_t}$ and move to another location in the scene, rendering vision as an inherently active, embodied process. The corresponding generative model is shown in **Figure 1B**, which is simplified from Parr et al. (2021), in the sense that Parr et al. (2021) also considers recursive definitions of entities, i.e., objects can again be defined as their constituent parts, and adopts a more fine grained factorization, e.g. also taking into account eye direction as separate factors.

Similar generative models can be used for learning machine vision using pixel observations (Eslami et al., 2018; Van de Maele et al., 2021b). In this case, the system is trained to make inferences about the scene $\mathbf{s}$, given images $\mathbf{o}_t$ and corresponding absolute viewpoints $\mathbf{v}_t$. This requires massive datasets containing many views of a large variety of scenes with a number of constituent objects, typically limited to primitive shapes and colors. However, this becomes unfeasible in the real world, where the variety of objects and their arrangement in scenes yields a combinatorial explosion, and where an accurate, absolute viewpoint of the camera is often missing. Also, developmental psychology suggests that toddlers don't learn from scanning scenes, but rather focus on a single dominating object that is close to the sensors (Smith et al., 2010).

Therefore, we propose a different generative model, which is more object-centric as opposed to scene-centric. We draw inspiration from the Thousand Brains Theory of Intelligence, focused on the computational principles of the neocortex (Hawkins et al., 2019). First, we subscribe to the principle of a repetitive functional unit, i.e., a cortical column, which have basic similarity of internal design and operation (Mountcastle, 1997). Second, each such functional unit learns a model of complex objects (Hawkins et al., 2017), inferring both "what" the object is as well as "where" it is located. We model a single repetitive unit to have both the "what" and "where" information streams, this in contrast to the brain anatomy where the ventral and dorsal stream are present in separate physical areas, resulting in separate cortical columns for this function (Hawkins et al., 2019). Additionally, our model only considers a single object per functional unit rather than the numerous models a cortical column in the brain can contain.

Third, instead of inferring both the observer's as well as the object's poses in a global reference frame, each model learns a representation in an object-centric reference frame (Hawkins et al., 2019). Again, the agent is enactive and can move around, but instead of changing an absolute location and/or head direction, actions are now encoded as relative displacements with respect to the object at hand. This is depicted in **Figure 2A**: at timestep $t$, the observer captures an observation $\mathbf{o}_t$ of a certain object with identity $\mathbf{i}$, at a certain pose $\mathbf{p}_t$ relative to the object. The observer can move around by executing action $\mathbf{a}_t$, which changes the relative viewpoint to $\mathbf{p}_{t+1}$.

We can formalize such an object-centric generative model as a Bayesian network, displayed in **Figure 2B**. We assume the agent focuses on a single object with identity $\mathbf{i}$, and can sample different poses $\mathbf{p}_t$ by moving around by taking actions $\mathbf{a}_t$. At each timestep $t$, the object identity $\mathbf{i}$ and current pose $\mathbf{p}_t$ yield the observation $\mathbf{o}_t$. The generative model up to the current timestep $t$ can then be factorized as:

$$P(\mathbf{i}, \mathbf{p}_{0:t}, \mathbf{o}_{0:t}, \mathbf{a}_{0:t-1}) = P(\mathbf{i})P(\mathbf{p}_0)\prod_{k=1}^{t}\underbrace{P(\mathbf{p}_k|\mathbf{p}_{k-1}, \mathbf{a}_{k-1})}_{\text{Transition model}}\underbrace{P(\mathbf{o}_k|\mathbf{p}_k, \mathbf{i})}_{\text{Likelihood model}}P(\mathbf{a}_{k-1}). \quad (1)$$

The generative model hence consists of a transition model, which models how an action moves the agent to a new poses, a likelihood model that predicts the observation of an object with a given identity viewed from a given pose, and prior distributions over identity, initial pose and actions.

Crucially, we will instantiate and learn such a separate model for each and every object type. The identity variable $\mathbf{i}$ then becomes a Bernoulli variable whether or not the object at hand belongs to the object type this particular model is representing. This is interesting from a computational perspective, as it allows to train each model on a confined dataset consisting of mainly views of a single object, which improves sample efficiency, and to instantiate a new model when a new object type is "discovered", enabling continual learning without catastrophic forgetting. To infer the object identity, we aggregate the outputs of the different models as having them casting a "vote."

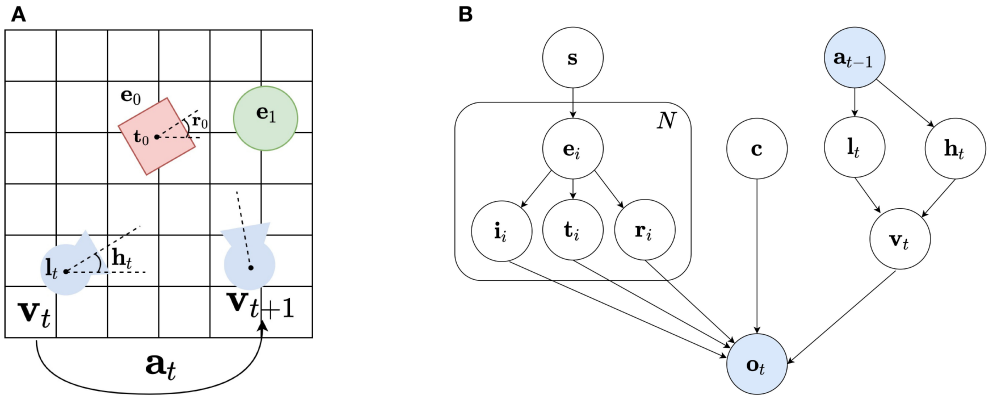



**FIGURE 1 | (A)** An observer's view of the world is determined by its location $\mathbf{l}_t$ and head direction $\mathbf{h}_t$ at a timestep $t$, and the objects $\mathbf{e}_i$ in the scene, their identity $\mathbf{i}_i$, and translation $\mathbf{t}_i$ and rotation $\mathbf{r}_i$ in the world coordinate frame. The observer can take action $\mathbf{a}_t$ to move to another location. **(B)** A generative model of vision, simplified from Parr et al. (2021): starting from a scene $\mathbf{s}$, we predict the objects or entities $\mathbf{e}_i$ one might encounter, their identity $\mathbf{i}_i$ and their placement in an allocentric reference frame defined by a translation $\mathbf{t}_i$ and rotation $\mathbf{r}_i$. Together with the context $\mathbf{c}$, i.e., the lighting conditions, and the viewpoint $\mathbf{v}_t$ of the observer the observation $\mathbf{o}_t$ is generated. Furthermore, the observer can change its viewpoint $\mathbf{v}_t$ by taking actions $\mathbf{a}_t$ that move its location $\mathbf{l}_t$ and/or head direction $\mathbf{h}_t$. Both actions and observations are observed variables shown in blue, whereas the others are unobserved and shown in white.

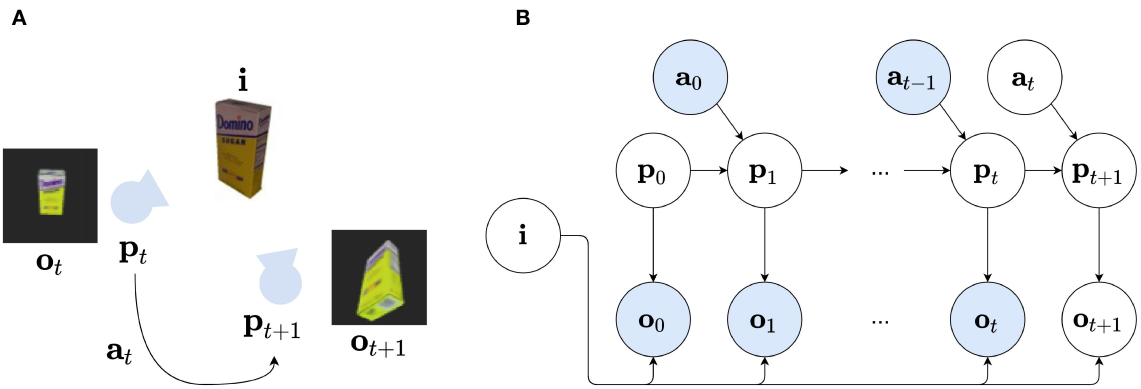



**FIGURE 2 | (A)** Visual representation of the environment in which an object with identity **i** (in this case: sugar box) can be observed by a camera at a pose $\mathbf{p}_t$, relative to the object. The agent can transform this viewpoint, provided it performs action $\mathbf{a}_{t+1}$ to go to pose $\mathbf{p}_{t+1}$. At each pose, an observation $\mathbf{o}_t$ is perceived. **(B)** The Bayesian Network describing the generative model of the agent. The variable $i$ represents the identity of the observed object, $\mathbf{p}_t$ represents the latent representation of the camera pose at timestep $t$. The variable $\mathbf{o}_t$ represents the sensory observation and is dependent on the identity $i$ and pose variable $\mathbf{p}_t$. The current camera pose $\mathbf{p}_t$ is dependent on the previous pose $\mathbf{p}_{t-1}$ and action $\mathbf{a}_{t-1}$ of the agent. Again observed variables are shown in blue, while unobserved variables are shown in white.

In what follows, we derive the (expected) free energy functional to infer actions for the agent to engage in active inference, and to update the model in doing so. Next, in Section 2.3, we provide more details on the actual parameterization of the model, the training mechanism and the voting scheme.

## 2.2. Active Inference

Active inference is a theoretical framework to describe the behavior of intelligent agents in dynamic environments. This theory postulates that all intelligent beings entail a generative model of the world, and act and learn in order to minimize an upper bound on the negative log evidence of their observations, i.e., free energy (Friston et al., 2016).

In order to infer beliefs about the unobserved variables, an agent needs to "invert" the generative model and calculate the posterior, which is in general intractable. Therefore, the agent resorts to variational inference, and approximates the true posterior by some tractable, approximate posterior distribution. In our case, we use an approximate posterior $Q(\mathbf{i}, \mathbf{p}_{0:t}|\mathbf{o}_{0:t})$ that factorizes as follows:

$$Q(\mathbf{i}, \mathbf{p}_{0:t}|\mathbf{o}_{0:t}) = Q(\mathbf{i}|\mathbf{o}_{0:t}) \prod_{k=0}^{t} Q(\mathbf{p}_k|\mathbf{i}, \mathbf{o}_k). \tag{2}$$

The variational free energy $F$ is a quantity to describe Bayesian surprise, i.e., how much the approximate posterior and the true

joint distribution differ. Given the generative model defined in Equation 1, the variational free energy $F$ is then defined as:

$$
\begin{aligned}
F &= \mathbb{E}_{Q(\mathbf{i},\mathbf{p}_{0:t})}[\log Q(\mathbf{i},\mathbf{p}_{0:t}|\mathbf{o}_{0:t}) - \log P(\mathbf{i},\mathbf{p}_{0:t},\mathbf{o}_{0:t},\mathbf{a}_{0:t-1})] \\
&= \underbrace{D_{KL}[Q(\mathbf{i}|\mathbf{o}_{0:t})||P(\mathbf{i})] + \sum_t D_{KL}[Q(\mathbf{p}_t|\mathbf{i},\mathbf{o}_t)||P(\mathbf{p}_t|\mathbf{p}_{t-1},\mathbf{a}_{t-1})]}_{\text{complexity}} \\
&\quad \underbrace{- \sum_t \mathbb{E}_{Q(\mathbf{i},\mathbf{p}_{0:t})}[\log P(\mathbf{o}_t|\mathbf{p}_t,\mathbf{i})]}_{\text{accuracy}}
\end{aligned} \tag{3}
$$

Hence, minimizing free energy entails maximizing model accuracy, while minimizing the model complexity, i.e., KL divergence between the approximate posterior and prior distributions. Also note that this is equivalent to maximizing the Evidence Lower Bound (ELBO) as used in variational autoencoders (Kingma and Welling, 2014; Rezende et al., 2014).

Crucially, in active inference, agents minimize the free energy not only by updating their internal model, but also by performing actions that they believe will minimize free energy in the future. However, future observations are of course not yet available. Therefore, the agent relies on its generative model to acquire expected observations over future states, and uses these to compute the expected free energy $G$ for an action $\mathbf{a}_t$:

$$
\begin{aligned}
G(\mathbf{a}_t) &= \mathbb{E}_{Q(\mathbf{i},\mathbf{p}_{0:t+1},\mathbf{o}_{t+1})}[\log Q(\mathbf{i},\mathbf{p}_{0:t+1}|\mathbf{o}_{0:t},\mathbf{a}_t) \\
&\quad - \log P(\mathbf{o}_{0:t+1},\mathbf{a}_{0:t-1},\mathbf{p}_{0:t+1},\mathbf{i}|\mathbf{a}_t)] \\
&\approx \underbrace{- \mathbb{E}_{Q(\mathbf{o}_{t+1})}[\log P(\mathbf{o}_{t+1})]}_{\text{instrumental value}} \\
&\quad \underbrace{- \mathbb{E}_{Q(\mathbf{i},\mathbf{p}_{0:t+1},\mathbf{o}_{t+1})}[\log Q(\mathbf{i}|\mathbf{o}_{0:t+1},\mathbf{a}_t) - \log Q(\mathbf{i}|\mathbf{o}_{0:t},\mathbf{a}_t)]}_{\text{info gain on object identity}} \\
&\quad \underbrace{- \mathbb{E}_{Q(\mathbf{i},\mathbf{p}_{0:t+1},\mathbf{o}_{t+1})}[\log Q(\mathbf{p}_{0:t+1}|\mathbf{i},\mathbf{o}_{0:t+1},\mathbf{a}_t) - \log Q(\mathbf{p}_{0:t+1}|\mathbf{i},\mathbf{o}_{0:t},\mathbf{a}_t)]}_{\text{info gain on object pose}}
\end{aligned} \tag{4}
$$

Here, we make two assumptions. First, we assume that the prior $P(\mathbf{o}_{0:t+1}|\mathbf{a}_t) \approx P(\mathbf{o}_{t+1})$. In active inference, the agent is assumed to have prior expectations about preferred future observations (Friston et al., 2016). Because this is a prior expectation, we can leave out the conditioning on action, and it only applies on future observations. Second, we assume that the bound on the evidence is tight, and hence that the approximate posterior distributions can be used in lieu of the true posteriors, i.e., $P(\mathbf{i}|\mathbf{o}_{0:t+1},\mathbf{a}_t) \approx Q(\mathbf{i}|\mathbf{o}_{0:t+1},\mathbf{a}_t)$ and $P(\mathbf{p}_{0:t+1}|\mathbf{i},\mathbf{o}_{0:t+1},\mathbf{a}_t) \approx Q(\mathbf{p}_{0:t+1}|\mathbf{i},\mathbf{o}_{0:t+1},\mathbf{a}_t)$.

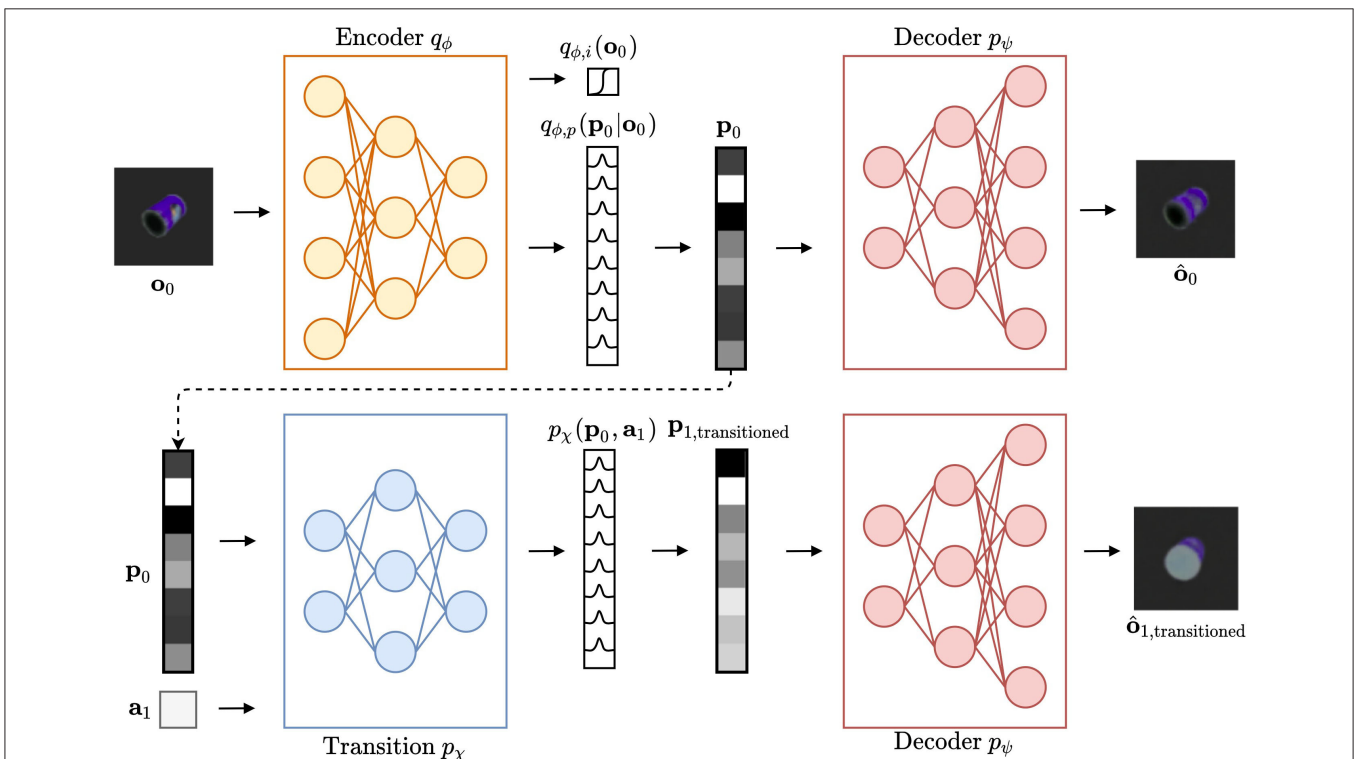



**FIGURE 3 |** A Cortical Column Network (modeling a master chef can). In the top left, an observation $\mathbf{o}_0$ is provided to the encoder model $q_\phi$. This model predicts the distribution over identity as Bernoulli variable to be either belonging to the dedicated object category (i.e., master chef can) or not. Secondly, a distribution over the latent pose variable $\mathbf{p}_0$ is predicted. A sample $\mathbf{p}_0$ from this distribution is then decoded through the decoder $p_\psi$ and provides the reconstruction $\hat{\mathbf{o}}_0$. Using the transition model $p_\chi$, this pose sample is transitioned into a belief over the latent pose variable $\mathbf{p}_{1,\text{transitioned}}$, given action $\mathbf{a}_1$. A sample from this new belief over $\mathbf{p}_{1,\text{transitioned}}$ is also decoded into an expected view $\hat{\mathbf{o}}_{1,\text{transitioned}}$.

The result can be decomposed into three terms. The first term is the instrumental value, which values future outcomes that have a high probability under the prior distribution over preferred outcomes. Intuitively, this will yield a high value for expected observations that are similar to the preferred observation. The second term is an epistemic term that values information gain on the object identity. This means that it will result in higher values for actions that will provide more information, i.e., the expected difference between prior and posterior distributions is large. The third term is also an epistemic term that values information gain on inferring the agent's pose relative to the object. This is similar to the second term, but this time in terms of the pose latent.

## 2.3. Cortical Column Networks

In order to engage in active inference, an implementation of the generative model is needed. We choose to model the vision system as the generative model defined in Section 2.1. We use the factorization shown in Equation (1). The priors over identity, initial pose and actions are constant and are therefore not explicitly modeled. The posterior distributions of the likelihood model is defined as the distribution over the observation, when the latent variables describing identity and pose are provided. The transition model represents the relation between the pose latent at the next timestep, provided with the pose latent at the current timestep and the taken action. Finally, we amortize the inference process that infers the latent variables describing identity and pose, given an observation by an encoder model. We call the combination of a likelihood model, transition model, and encoder model for a single object category a Cortical Column Network (CCN) for this object category. In this context, amortization simply means learning a mapping from sensory input to the sufficient statistics of an approximate posterior, with a known functional form. Knowing the functional form of the posterior means the free energy objective functionals are well defined, enabling the application of standard optimization techniques (in this case Adam Kingma and Ba, 2015). This enables a generic optimization of belief distributions that underwrite active inference (Dayan et al., 1995), and can be thought of as learning to infer.

For high-dimensional data, such as pixel-based observations, designing a mapping to a latent distribution is infeasible by hand. We thus resort to deep learning to learn the likelihood and transition models directly from observation data. Additionally, we amortize the inference process and learn the encoder model jointly, similar to the approach applied in variational autoencoders (Kingma and Welling, 2014; Rezende et al., 2014).

### 2.3.1. Model

We propose the Cortical Column Network (CCN) as basic building block of our architecture. Drawing inspiration from the Thousand Brains Theory (Hawkins et al., 2017), which promotes the modularity of cortical columns in the brain that learn predictive models of observed objects, we instantiate a separate CCN for each object type or identity. This results in a dedicated CCN for each known object type, and can be scaled to more objects by adding more CCNs. A CCN consists of three neural networks: an encoder $q_\phi$, a decoder $p_\psi$, and a transition model $p_\chi$, which parameterize the approximate posterior, likelihood model and transition model introduced in Equations (1) and (2). The encoder $q_\phi$ has two heads that map a pixel-based observation to both a pose latent space $\mathbf{p}$, which is modeled as a Normal distribution with a diagonal covariance matrix, and an identity latent space $i$, modeled as a Bernoulli variable. The decoder $p_\psi$ learns the mapping from the pose latent $\mathbf{p}$ to a distribution over the observation $\mathbf{o}$, which is modeled as a Normal distribution with fixed variance $\mathcal{N}(\hat{\mathbf{o}}, I)$. The transition model $p_\chi$ learns to transform a sample from the pose latent $\mathbf{p}$ to a belief over the transitioned pose in latent space, also modeled as a Normal distribution with diagonal covariance matrix.

The information flow of a single CCN is shown in **Figure 3**. A single CCN is dedicated to model a single object type, in this case a master chef can. An observation $\mathbf{o}_0$ is fed into the encoder $q_\phi$, as depicted in the top left corner. The belief over the identity of observation $\mathbf{o}_0$ is represented as a Bernoulli variable marking whether or not the observation belongs to the CCN object category. The encoder also outputs a distribution for the pose latent, from which samples can be decoded into expected observations using decoder $p_\psi$, as shown in the top right of the figure. Finally, the bottom row illustrates the transition model $p_\chi$, which computes a belief over the pose latent $\mathbf{p}_1$ after taking an action $\mathbf{a}_1$, at current pose latent $\mathbf{p}_0$. Again, the decoder model $p_\psi$ can be used to estimate observation $\hat{\mathbf{o}}_1$ after action $\mathbf{a}_1$. This gives the CCN the ability to imagine "what would this object look like from here," and to infer the best action, e.g. that minimizes the expected free energy (Equation 4). Once an action is selected, the agent moves to a new pose, obtains a novel observation $\mathbf{o}_1$, and the process repeats.

### 2.3.2. Optimization

The encoder, decoder and transition neural networks for a single object are optimized in an end-to-end manner from pixel-based observations. For each object, we create a dataset $\mathcal{D}_i$ from which one can sample triplets $(\mathbf{o}_0, \mathbf{a}_1, \mathbf{o}_1)$, i.e., two images $\mathbf{o}_0$ and $\mathbf{o}_1$ together with action $\mathbf{a}_1$ which is the relative transform to move the camera from the initial to the next viewpoint. All viewpoints are collected such that the target object is centered in view.

The overall train procedure is given in **Algorithm 1**. When training the CCN for object $i$, each iteration we sample a triplet $(\mathbf{o}_0, \mathbf{a}_1, \mathbf{o}_1)$ from $\mathcal{D}_i$, as well as an observation $\mathbf{o}_{\text{negative}}$ of a random other dataset $\mathcal{D}_{j\neq i}$. We forward all observations through the encoder model, and reconstruct $\hat{\mathbf{o}}_0$, $\hat{\mathbf{o}}_1$ from the pose latents $\mathbf{p}_0$ and $\mathbf{p}_1$, as well as $\hat{\mathbf{o}}_{1,\text{transitioned}}$ after transitioning from $p_\chi(\mathbf{p}_0, \mathbf{a}_1)$. To minimize the variational free energy as defined in Equation (3), our loss function becomes:

$$\mathcal{L}_{\text{FE}} = \underbrace{||\hat{\mathbf{o}}_0 - \mathbf{o}_0||_2 + ||\hat{\mathbf{o}}_1 - \mathbf{o}_1||_2 + ||\hat{\mathbf{o}}_{1,\text{transitioned}} - \mathbf{o}_1||_2}_{\mathcal{L}_{\text{reconstruction}}}$$
$$+ \underbrace{D_{\text{KL}}[q_{\phi,\mathbf{p}}(\mathbf{o}_1)||p_\chi(\mathbf{p}_0, \mathbf{a}_1)]}_{\mathcal{L}_{\text{complexity}}}$$
$$+ \underbrace{\text{BCE}(q_{\phi,i}(\mathbf{o}_0), 1) + \text{BCE}(q_{\phi,i}(\mathbf{o}_1), 1) + \text{BCE}(q_{\phi,i}(\mathbf{o}_{\text{negative}}), 0)}_{\mathcal{L}_{\text{classification}}} \tag{5}$$

Here, we represent the likelihood model as an isotropic Gaussian on the reconstructed pixels $\hat{\mathbf{o}}$, which yields a mean squared reconstruction loss for the accuracy term in Equation 3, resulting in the $\mathcal{L}_{\text{reconstruction}}$ term of the loss in Equation (4). $\mathcal{L}_{\text{complexity}}$ in Equation (4) exactly represents the complexity term for the poses as a KL divergence term between the encoded pose distribution on the one hand, and the predicted transitioned pose distribution on the other hand. For object identity, we assume a uniform prior $P(\mathbf{i})$ in Equation (3), which results in a binary cross entropy (BCE) loss term for each CCN, and we use the other object sample $\mathbf{o}_{\text{negative}}$ to contrast. These terms form the $\mathcal{L}_{\text{classification}}$ of the loss in Equation (4). For further details on the training procedure, we refer to the implementation details in Section 3.1. Note that the distribution over the latent pose variable is modeled as a Gaussian distribution with a diagonal covariance matrix for which the parameters are learned through the optimization process. Hence, these latent dimensions do not reflect the translation and orientation parameters of an absolute pose in an Euclidean reference frame, but encode the pose in an abstract, object-local reference frame.

**Algorithm 1** : CCN training.

**for** iteration $= 1, 2, \ldots$ **do**
  $(\mathbf{o}_0, \mathbf{a}_1, \mathbf{o}_1) \sim \mathcal{D}_i$ ▷ Sample observation-action pairs from the dataset of object identity i
  $\mathbf{o}_{\text{negative}} \sim \mathcal{D}_{j \neq i}$ ▷ Sample negative anchor
  $\mathbf{p}_0, i_0 \sim q_\phi(\mathbf{o}_0)$ ▷ Encode the observations and sample a pose and identity latent
  $\mathbf{p}_1, i_1 \sim q_\phi(\mathbf{o}_1)$
  $\mathbf{p}_{1,\text{transitioned}} \sim p_\chi(\mathbf{p}_0, \mathbf{a}_1)$ ▷ Transition the pose latent
  $\hat{\mathbf{o}}_0 \leftarrow p_\psi(\mathbf{p}_0)$ ▷ Reconstruct samples
  $\hat{\mathbf{o}}_1 \leftarrow p_\psi(\mathbf{p}_1)$
  $\hat{\mathbf{o}}_{1,\text{transitioned}} \leftarrow p_\psi(\mathbf{p}_{1,\text{transitioned}})$
  ▷ Compute the loss terms
  $\mathcal{L}_{\text{reconstruction}} \leftarrow ||\hat{\mathbf{o}}_0 - \mathbf{o}_0||_2 + ||\hat{\mathbf{o}}_1 - \mathbf{o}_1||_2 + ||\hat{\mathbf{o}}_{1,\text{transitioned}} - \mathbf{o}_1||_2$
  $\mathcal{L}_{\text{classification}} \leftarrow \text{BCE}(q_{\phi,i}(\mathbf{o}_0), 1) + \text{BCE}(q_{\phi,i}(\mathbf{o}_1), 1) + \text{BCE}(q_{\phi,i}(\mathbf{o}_{\text{negative}}), 0)$
  $\mathcal{L}_{\text{complexity}} \leftarrow D_{\text{KL}}[q_{\phi,p}(\mathbf{o}_1)||p_\chi(\mathbf{p}_0, \mathbf{a}_1)]$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{classification}} + \mathcal{L}_{\text{complexity}}$
  $\phi, \chi, \psi \leftarrow \text{Adam}(\mathcal{L})$ ▷ Update parameters

### 2.3.3. Voting Over Object Identity

After training a CCN for each of the $N$ known objects, our aim is to infer the object identity $Q(\mathbf{i}|\mathbf{o}_{0:t})$, as a categorical distribution with $N + 1$ categories, one for each object type and an "other" category. To this end, we use a Dirichlet distribution with concentration parameters $\alpha_{0:N}$ as conjugate prior for the categorical variable. At each timestep $t$, the concentration parameters are updated as follows (Smith et al., 2022):

$$\begin{cases} \alpha_{i,t} & = \alpha_{i,t-1} + \eta \cdot q_{\phi,i}(\mathbf{o}_t)\text{, for } i < N \\ \alpha_{N,t} & = \alpha_{N,t-1} + 0.5 \end{cases} \tag{6}$$

We initialize $\alpha_{i,0}$ as a constant vector with values 0.1. This can be interpreted as the voting mechanism from the Thousand Brains theory (Hawkins et al., 2017), where each CCN casts a vote on whether the object in view belongs to the category it was trained on. Over time, the different votes are aggregated as collecting evidence for the different object categories. When an unambiguous view is rendered from a known object, only a single CCN, i.e., the one trained on that object category, will be active and cast a vote. However, in the case the object category cannot be distinguished from an observation, i.e., the top of a cylindrical object could be both a master chef can or a chips can, multiple votes will be cast on the different possible categories. In this case, the embodied agent can query additional views, in particular views that will provide information gain about object identity and as such minimizing the expected free energy defined in Equation 4.

In case of an unknown object, ideally none of the CCNs is active. Therefore, we add a fixed vote of 0.5 for the "other" category, which will prevail when none of the CCNs is consistently active over time. However, in practice, we find that unkown objects behave as out-of-distribution data for each individual CCN, and the predictions from the learned model are therefore unreliable. To mitigate this inherent limitation of deep neural networks, we propose an additional likelihood-based scheme for detecting the "other" category. Concretely, we look at the reconstruction error of the likelihood model to assess whether the CCN is in effect correctly modeling the object at hand. When the reconstruction error exceeds an object-specific threshold, the votes cast by the CCNs are ignored, i.e., $\eta = 0$, and only a vote of 0.5 is cast for the "other" category.

Moreover, instead of calculating the total mean squared error, we use a scaled reconstruction error. As scale factor, we choose the reciprocal of the amount of pixels in the intersection between the foreground masks of the prediction and the observation. The foreground masks are obtained by thresholding the fixed background color used in the renderings. This forces the original observation and the reconstruction to have high overlap, and increases the weight of foreground pixels for small objects.

When multiple timesteps are considered, the likelihood based threshold also considers the transition with respect to the previous observation. Concretely, when executing an action, we predict the new observation by first inferring the new pose given the previous pose and action, and reconstructing that one. Again,

in order for the vote to be valid, the CCN must now have a scaled reconstruction error smaller than the thresholds for both cases.

In the case of embodied agent, the action selection process is driven through the minimization of expected free energy $G$. To infer the object identity, the prevalent term in the expected free energy $G$ is the information gain term on object identity. The agent then chooses the action as follows:

$$\mathbf{a}_{t+1} = \underset{\mathbf{a}_{t+1}}{\operatorname{argmin}} -\mathbb{E}_{Q(\mathbf{i},\mathbf{p}_{0:t+1},\mathbf{o}_{t+1})}[\log Q(\mathbf{i}|\mathbf{o}_{0:t+1},\mathbf{a}_t) - \log Q(\mathbf{i}|\mathbf{o}_{0:t},\mathbf{a}_t)] \tag{7}$$

In practice, we use a Monte Carlo approximation where we evaluate this term for a number of randomly sampled actions, and select the best one. Similarly, the expectation is approximated by sampling from our models.

#### 2.3.4. Moving Toward a Preferred Observation

Once the agent has inferred the object class and its pose with respect to the object, it can also use the model to infer actions that bring the agent toward a preferred observation $\mathbf{o}_{\text{preferred}}$. This can be useful in use cases where the agent needs to inspect a particular aspect of a certain object more closely, or when the agent needs to manipulate the object and is provided with a (demonstration of a) grasp pose.

To infer the action that brings the agent toward a preferred observation, we can again evaluate the expected free energy $G$. In this case, we assume the agent already correctly inferred the object identity and pose, i.e., the information gain on these variables is low, and the expected free energy $G$ boils down to maximizing the instrumental value in Equation (4), i.e., the expected error between the predicted and preferred observation. As our likelihood model in pixel-space does not necessarily reflect the perceptual difference between two images (Zhang et al., 2012), we match instead the likelihood in the pose latent space. We do this by first determining the preferred pose distribution $P(\mathbf{p}_{t+1})$ by encoding the preferred observation $\mathbf{o}_{\text{preferred}}$, and then minimizing the expected free energy with respect to the actions to match this preferred distribution, essentially computing:

$$\mathbf{a}_{t+1} = \underset{\mathbf{a}_{t+1}}{\operatorname{argmin}}\, \mathbb{E}_{Q(\mathbf{p}_{t+1}|\mathbf{p}_t,\mathbf{a}_{t+1})}[-\log(P(\mathbf{p}_{t+1}))] \tag{8}$$

Again using a Monte Carlo approximation, we first sample random actions, evaluate the expected free energy for all these actions with respect to the preferred pose distribution, and select the action with the lowest expected free energy. The preferred pose distribution is computed by encoding the preferred observation $\mathbf{o}_{\text{preferred}}$ using the encoder model $q_{\phi,\mathbf{p}}$, whereas the expected pose distribution is acquired by transitioning the current pose latent $\mathbf{p}_t$ to an expected future pose latent using the transition model $p_\chi$.

**TABLE 1 |** Ranges from which the absolute viewpoints are sampled in spherical coordinates in the dataset creation process.

| Variable | min | max |
|---|---|---|
| azimuth | 0 | $2\pi$ |
| elevation | $-\frac{\pi}{2}$ | $\frac{\pi}{2}$ |
| radius | 0.10m | 0.55m |
| $\theta$ | 0 | $2\pi$ |

## 3. RESULTS

In this section, we conduct and analyze a number of experiments to evaluate our proposed approach. First we explicate the experimental setup, dataset creation, model parameterization, and training details. In a series of experiments the following research questions are addressed:

- Can a collection of CCNs be used for object classification?
- Can the ensemble of CCNs be used for detecting which object categories are out of distribution, essentially quantifying what the model does not know?
- Does embodiment improve classification accuracy as the agent can resolve ambiguity using multiple observations?
- Can a CCN for a given object category be used for object pose estimation?

### 3.1. Experimental Setup

To train our ensemble of CCNs, a dataset of different objects is required. To this end we select a subset of 33 objects from the YCB dataset (Calli et al., 2015), for which high quality triangular meshes were readily available. This set of objects is split in a known and unknown category, consisting of 26 and 7 objects respectively. For a full list of the used objects, the reader is referred to the **Supplementary Materials**.

For each object category of the known category, we create a dataset by rendering object meshes from this object on a uniform background. The camera poses are sampled randomly from a uniform distribution in spherical coordinates, for which the ranges are provided in **Table 1**. The orientation is then determined as the orientation to point the camera to the center of the object's bounding box, and randomly rotated with angle $\theta$ around the axis pointing to the object. For each object, a dataset of 10000 views is created, for which 90% is used as train data, 5% as validation data and 5% for testing.

We base our encoder and decoder model on the variational autoencoder architecture used in Ha and Schmidhuber (2018), where an image is first processed through a convolutional pipeline, after which a linear layer is used to transform the extracted information into the parameters of a Gaussian distribution with a diagonal covariance matrix. The decoder is the inverse of this process, where the embedding is expanded into the spatial dimensions. This result is then upscaled through a deconvolution pipeline into an expected observation. For the transition model, we simply use a multilayer perceptron network.

The encoder model $q_\phi$ is instantiated as a convolutional neural network that first processes a 64 by 64 RGB image with

4 convolutional layers. Each layer has a 4x4 kernel and uses a stride of 2. The layers output tensors with 8, 16, 32, and 64 channels, respectively after which they are activated through a LeakyReLU activation function with a negative slope of 0.01. The resulting representation is flattened to a 256-dimensional vector after which it is processed by two separate heads, or in other words, separate linear layers. The classification head is a linear layer, followed by a sigmoid activation function that predicts the Bernoulli variable directly. The second head predicts the mean of the belief over the pose latent by a linear layer with 8 outputs, while the variance is predicted as the softplus of the output of a third linear layer with 8 outputs.

The decoder model $p_\psi$ is designed as the inverse of the encoder. The latent code is first expanded into a 64 dimensional vector using a linear layer, followed by a LeakyReLU (0.01 negative slope). The result is now reshaped into an image tensor that can be processed by convolutional layers. It is then processed by 2 transposed convolution layers with kernel size 6 and stride 2, after which it is fed through 2 transposed convolutions with kernel size 5 and stride 2. The output channels of these four layers are 64, 64, 32 and 16 and are followed by LeakyReLU activations with a negative slope of 0.01. Finally, a convolution layer with kernel size 1x1 and stride 1 is used to compress the channels into a 3-channel image, followed by a sigmoid to ensure the outputs are in the $[0, 1]$-range.

The transition model $p_\chi$ is parameterized as three linear layers that are followed by a LeakyReLU activation function with negative slope of 0.01. The first layer takes the concatenation of the pose latent code, the translation vector of the selected action and the orientation quaternion of the selected action as input, and transforms it to a 128 dimensional vector. The following two linear layers both have 256 outputs. This final output is then passed through 2 separate linear layers with 8 outputs, of which the first represents the mean of the transitioned belief and the second is passed through a softmax, which then represents the predicted variance of the belief over the transitioned pose.

The model is optimized in an end-to-end fashion using the Adam optimizer (Kingma and Ba, 2015) with learning rate $10^{-4}$ on the loss described in Equation (4). The separate terms in this loss function are scaled using Lagrangian multipliers (Rezende and Viola, 2018), which are inversely proportional to the gradient on the difference between the loss-term and a tolerance, to avoid posterior collapse. The multipliers for each term have an initial value and will be adapted within a specific range. The tolerances start at a fixed, low value and are updated every 500 steps. If the threshold is not reached, the tolerance is relaxed by multiplying it with a value of 1.10. This enforces the model to focus first on producing good reconstructions, and later optimize for classification and minimizing complexity. We also add a KL loss for all Gaussian outputs to standard normal to improve training stability. The values used in the optimization process are shown in **Table 2**.

**TABLE 2 |** Values used in the constrained optimization mechanism (Rezende and Viola, 2018), used for training a CCN.

| Parameter | Initial value | Range |
|---|---|---|
| $\lambda_{\text{reconstruction}}$ | 80 | [0, 100] |
| $\lambda_{\text{reconstruction_transition}}$ | 40 | [0, 100] |
| $\lambda_{\text{classification}}$ | 500 | [0, 1000] |
| reconstruction start tolerance | 10 | N/A |
| transitioned start reconstruction tolerance | 10 | N/A |
| classification start tolerance | 0.01 | N/A |
| tolerance adjust frequency | 500 | N/A |

*Reconstruction references direct reconstruction without transitioning the latent code. The transitioned reconstruction references the reconstruction loss on the transitioned latent, and the classification terms reference the binary cross entropy terms in the loss function from Equation (4).*

## 3.2. Classification

First, we investigate the classification performance of our ensemble model consisting of 26 CCNs. These CCNs are each trained on a single object category, while views from the other 25 categories are used as negative anchors. The 26 object categories are listed in the confusion matrix, shown in **Figure 4**. First, we evaluate the performance of classifying a single observation, followed by an experiment in which an embodied agent can query multiple observations sequentially.

### 3.2.1. Static Agent Classification

To investigate the classification performance of a static agent, we provide the agent with a single observation. We address whether an ensemble of CCNs can be used for accurate object classification. Additionally, we investigate to what extent our approach can accurately detect when an object is out of distribution, i.e., the object does not belong to a category previously seen by the agent during training.

For each object category, 100 samples are randomly sampled from the test for classification, and all unknown objects are clustered in an "other" category. As described in Section 2.3, each CCN votes for the known category it was trained on, provided that the reconstruction likelihood is within a predefined threshold. We empirically determine the threshold for each category by looking at the reconstruction errors of train-set observations, and scale the 95% quantile value by a factor 1.1, to remove outliers. This results in a high classification performance while still being able to detect more novel objects.

We show the confusion matrix for the static agent in **Figure 4**. An average classification accuracy of 86.71% is achieved. The confusion matrix shows that the main source of errors is due to the CCNs not being confident enough on the reconstruction and the "other" vote wins. We also see that in some cases there is some confusion between similar shaped objects, i.e., between "pudding box," "cracker box," and "gelatin box." We hypothesize (see Section 3.2.2) that querying more observations of the same object will adjust the vote for the correct object category, and after multiple observations the agent will resolve these issues. We qualitatively investigate these difficult samples, as is shown in **Figure 5**. This figure shows ambiguous observations that are incorrectly classified by the ensemble of CCNs. It can be observed that the reconstruction from both the (wrongly) chosen model and the correct model are very similar. For example for the

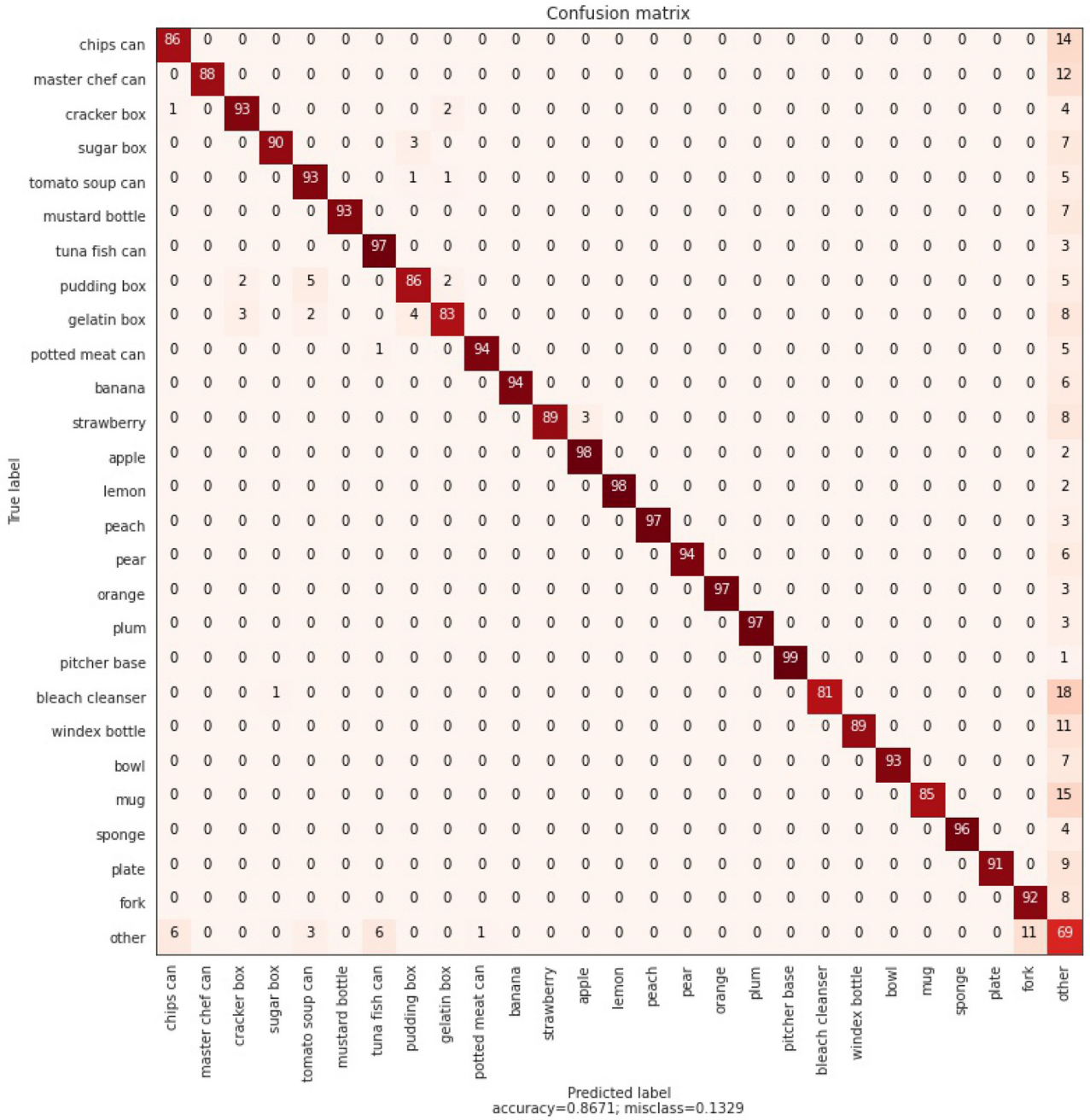



**FIGURE 4 |** Confusion matrix, using the max of a Dirichlet distribution and a likelihood based threshold over object beliefs as described in Section 2.3. 100 examples are classified for each class of the 26 known and 7 unknown objects. Overall, an average classification accuracy of 86.71% is achieved.

strawberry and the apple, a large red circle is reconstructed. It is thus difficult to accurately predict the object class.

**Figure 5** also shows the expected next viewpoint that would be encountered if an action minimizing expected free energy $G$ was performed. The latent code for both potential object categories is acquired through both the correct and incorrectly chosen transition models and an imagined view can be acquired using the respective decoders. Clearly, these selected observations are more easily distinguishable and thus enforce our hypothesis that embodiment will aid in the correct classification of three dimensional objects.

Alternatively, we could also train a single classifier using the same amount of parameters as the ensemble of CCNs which we expect to achieve similar classification accuracy to the ensemble of CCNs. However, due to the inherent nature of the model design, it would be unable to estimate when

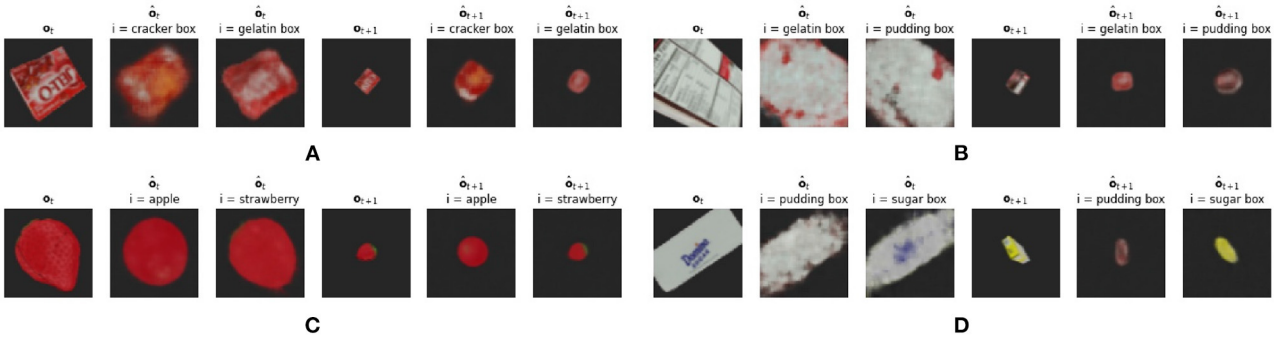



**FIGURE 5 |** Ambiguities found in the ensemble CCN classifier. From left to right: initial observation $\mathbf{o}_t$ provided to the agent. The reconstruction $\hat{\mathbf{o}}_t$ predicted using the expected model. The reconstruction $\hat{\mathbf{o}}_t$ predicted using the correct model. The chosen next observation $\mathbf{o}_{t+1}$ through the minimization of expected free energy $G$. The expected observation $\hat{\mathbf{o}}_{t+1}$ using the selected model. The expected observation $\hat{\mathbf{o}}_{t+1}$ using the correct CCN model. **(A)** The gelatin box. **(B)** The pudding box. **(C)** The strawberry. **(D)** The sugar box.

objects from an unseen category occur and this model would not have the flexibility to add new modules when novel objects are encountered.

### 3.2.2. Embodiment and Aggregating Votes

The previous section showed some clear disadvantages using a static agent: ambiguities can not be resolved, nor can information from previous observations be used to make a more confident, and more reliable decision. In the following experiment, we investigate whether classification accuracy improves as the agent is embodied and can actively query novel viewpoints. At each timestep, the agent randomly samples 100 action candidates and evaluates the expected free energy, i.e., to maximize the information gain on object identity as stated in Equation (4). This action is then executed and a novel observation is provided to the agent, which updates the Dirichlet concentration parameters and the process repeats.

In **Figure 6**, the classification accuracy of an active inference driven agent over time is shown for different datasets. When testing the agent only on the known classes (dashed line), the agent can immediately resolve ambiguities and the performance reaches 100% after two steps. When only considering objects from the 7 "other" categories (dotted line), the classification accuracy starts at a lower value of around 70% (as can also be seen in the confusion matrix in **Figure 4**), and over time reaches an accuracy of 85%. Finally, the red line shows the classification accuracy for all objects combined (26 known and 7 unknown objects). The performance rises from 87 to 97% after nine steps. The full confusion matrix for each different step can be found in the **Appendix**.

It can be observed that the accuracy for the known classes only increases. This is attributed to the Dirichlet information aggregation scheme. As more information is acquired, the votes and evidence for certain object categories becomes more overwhelming. In contrast, accuracy for the other category clearly gains information after a single timestep, but then fluctuates between 80 and 90%. As described in Section 2.3, the other category is mainly detected by the second reconstruction-based phase of the classification pipeline. This phase considers the current observation, and the transition given the previous observation, the window of information is thus two timesteps, and therefore no classification performance increase is found after more than two steps.

As a comparison baseline, we evaluated the embodiment using a random agent, i.e., the next viewpoint is randomly selected instead of using free energy minimization. The accuracy this random agent realizes, is indicated by the blue line in **Figure 6**. The performance of the random agent is on par with the active inference agent. We also observe that the ratio of informative views with respect to ambiguous views is high. Recall from the confusion matrix, the correct object identity can be inferred in over 80% in the first step of the (randomly) sampled views. Hence, it is to be expected that providing a random additional view provides the necessary information to get the correct classification, and the free energy agent has only a small margin to improve upon. We expect the gap between the free energy agent and a random agent to become larger in the case where more ambiguous viewpoints are present, as the free energy agent will avoid those as evidenced by **Figure 5**.

## 3.3. Pose Estimation

Next, we evaluate to what extent CCNs can be used for object pose estimation, given a desired view. First, we qualitatively evaluate the object pose estimation for different objects. At each timestep, the agent samples 1,000 random actions and calculates the instrumental term of expected free energy $G$ as described in Section 2.3.4. Again, the agent selects the action that minimizes the expected free energy and queries a new observation.

In **Figure 7**, we plot the input and target views, as well as the predicted viewpoints with the best (lowest) and worst (highest) expected free energy $G$ for master chef can, mustard bottle, strawberry and windex bottle. Below each observation the inferred or predicted latent code is shown. It is clear that the latent code is similar for matching the observations, while having the additional benefit that it does not suffer from the typical issues with MSE such as the scaling issue for pixel-wise errors.

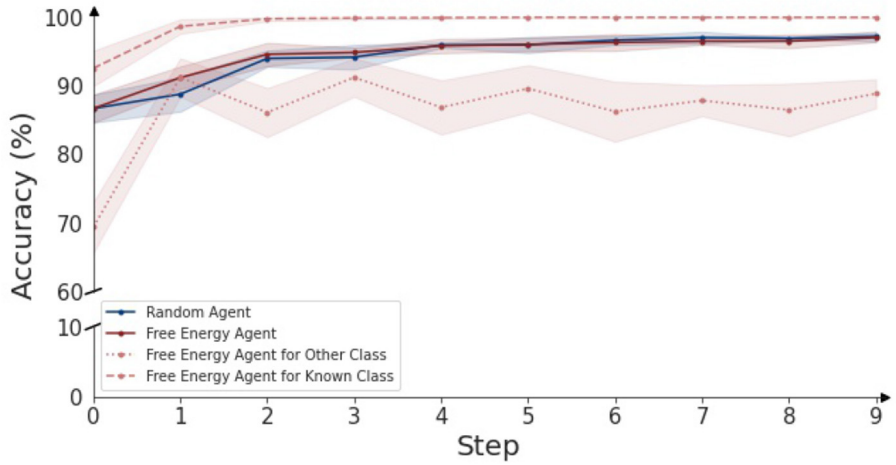



**FIGURE 6 |** Classification accuracy over time for an embodied agent, driven through the active inference paradigm. The agent is provided with different objects in random poses to classify, accuracy over a duration 10 steps is plotted. For each object category, 5 splits of 20 observations are classified and are used to visualize the 95% confidence bounds. The graph indicates classification accuracy over time for objects of the 26 known and 7 unknown objects. The red line represents the accuracy for the free energy agent, while the blue line represents the accuracy for the random agent. For the active inference agent, the distinction is made between the known and unknown objects: the dotted red line indicates the classification accuracy for objects of the 7 unknown object categories and the dashed line indicates the classification accuracy for the 26 known objects.

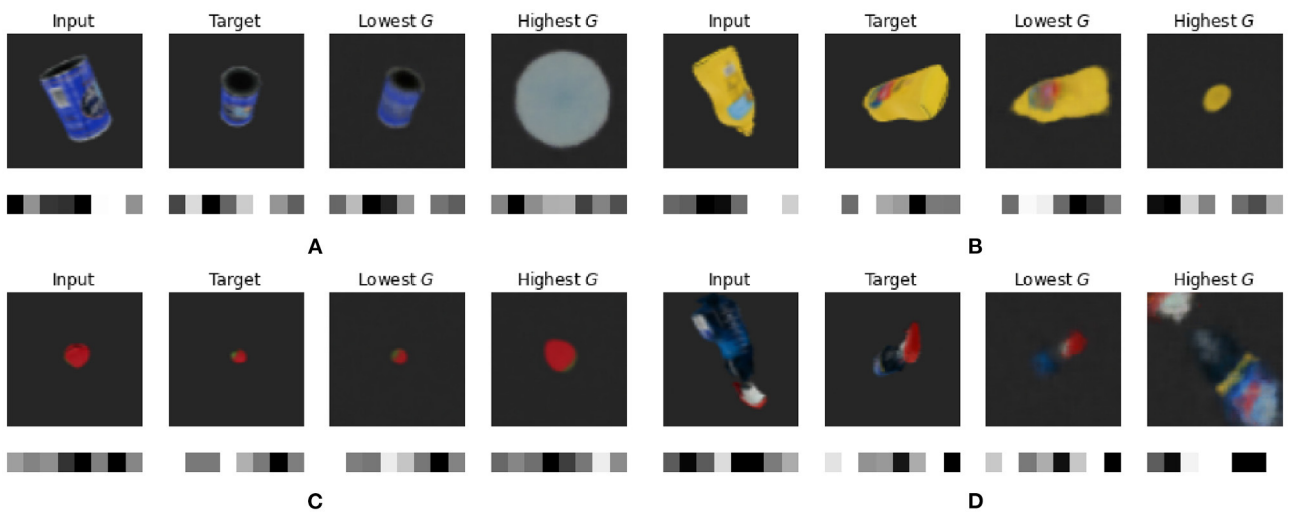



**FIGURE 7 |** Qualitative results for object pose estimation for **(A)** master chef can, **(B)** mustard bottle, **(C)** strawberry, and **(D)** windex bottle. The first column shows the input of the model with the mean latent code shown below. The second column shows the target observation along with its mean latent code. The third column shows the imagined observation and the transitioned latent code for the action with the lowest expected free energy $G$, while the final column shows the imagined observation and latent code for the action with the highest expected free energy $G$.

However, when we quantitatively evaluated the resulting poses, we noticed that the absolute pose error in Euclidean space was often way off, despite similar reconstructions. To further inspect this, we plot the expected free energy landscape for varying azimuth and elevation for the predicted target pose, as well as the initial, target and selected pose.

In **Figure 8**, we show heatmaps of expected free energy $G$ for 25 objects from the YCB dataset in the pose estimation scheme. The exact pose can be represented by four degrees of freedom: azimuth, elevation, radius and axis angle $\theta$. We vary two of these dimensions while keeping angle $\theta$ and radius fixed and plot the expected free energy landscape for the agent to reach a target observation, marked by the red cross. As indicated by the figure legend, the lightly colored areas are more desired by an active inference agent as they have a lower expected free energy. The red cross marks the preferred state of the agent, and the black dots show the initial observation and the point with lowest expected free energy.

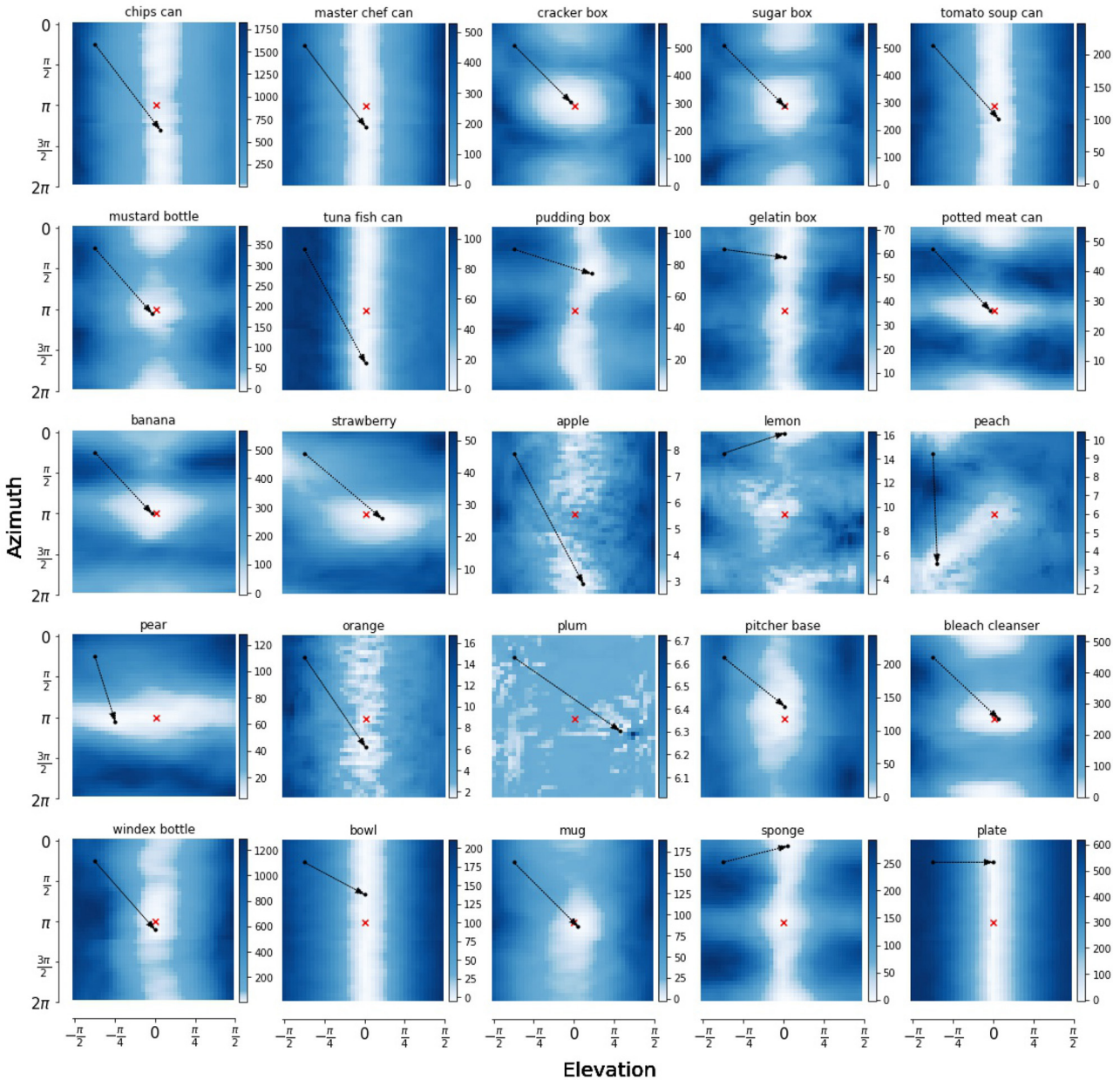



**FIGURE 8 |** Heatmaps of the expected free energy $G$ for reaching a target observation, marked by the red x. The pose parameters radius and angle $\theta$ around the z-axis of the camera are kept in a fixed position, while varying the azimuth and elevation angle over their full range in 40 linearly spaced points in each dimension. The initial point and the point with lowest expected free energy $G$ are marked by black dots. The arrow head points at the point with lowest free energy, starting at the pose of initial observation.

For some objects, such as banana and mug, there is a clear global minimum in the expected free energy landscape, and the pose estimation is quite accurate. However, for other objects, such as sugar box, mustard bottle and bleach cleanser, there are multiple local optima, or the landscape might even be invariant to the azimuth axis, as is the case for a lot of the can and box objects, the bowl and plate. These areas with low free energy are aliased areas, where the symmetry of the object surfaces. This shows how our model has actually learned various object symmetries, and learns to map different aliases with similar pixel observations onto the same point in pose latent space.

This can be viewed more clearly for imaginations generated while varying one dimension of these plots. **Figure 9** shows imaginations of a varying azimuth or elevations while keeping the three other dimensions fixed. In the heatmap of the

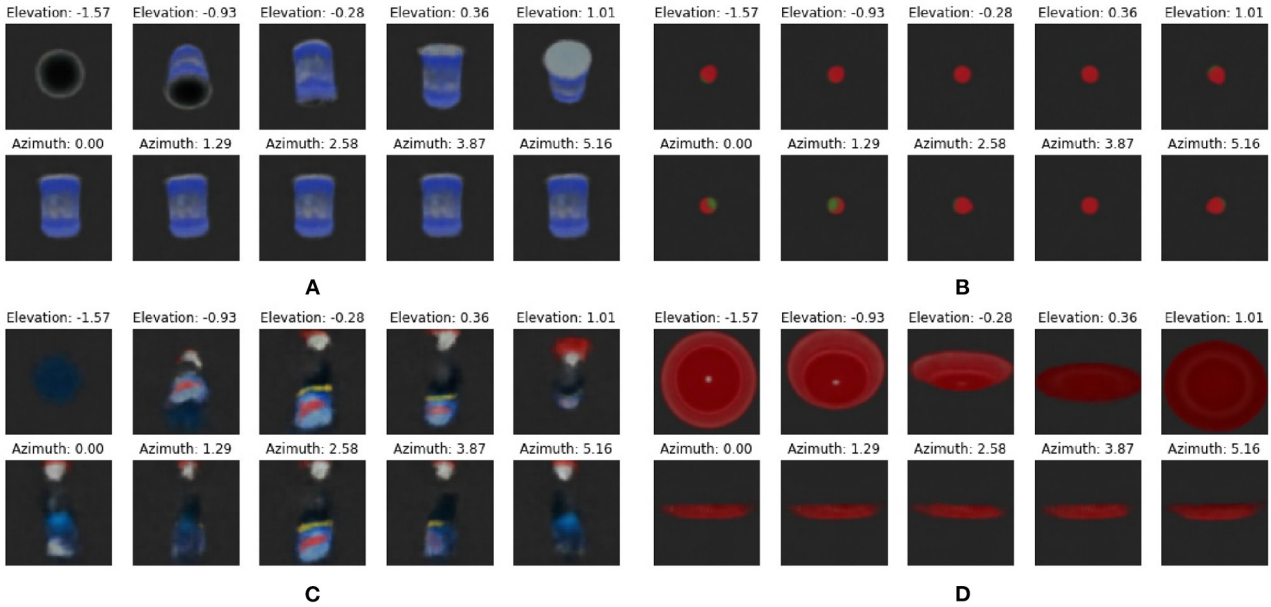



**FIGURE 9 |** Imaginations generated from CCN models of **(A)** master chef can, **(B)** strawberry, **(C)** windex bottle, and **(D)** plate. We vary a single dimension while keeping the other three at a fixed position. The top row shows a varying elevation and the bottom row shows a varying azimuth, within the full range as defined in Section 2.3.

master chef can, it can be observed that varying the azimuth results in the same expected free energy, while this differs for changing the elevation. **Figure 9A** shows this more clearly as all reconstructions of a straight can are identical. The model did not learn to reconstruct the exact contents of the can label, but blurs this as an average of the whole can. Similarly for **Figure 9D**, a different azimuth will show a horizontal plate, and no difference can be found. In contrast, the strawberry has a fairly localized expected free energy minima, which can be attributed to the position of the green on the strawberry head. For this reason, the different orientations can be differentiated. The same can be found for the windex bottle, where the objects inherent asymmetry results in a clearer loss landscape.

## 4. DISCUSSION

In this article, we proposed a method for learning object-centric representations in an unsupervised manner from pixel data. We draw inspiration from recent theories in neuroscience, in particular an active inference account of human vision (Parr et al., 2021) and the Thousand Brain Theory on intelligence (Hawkins et al., 2017). This leads to a modular architecture, where each model separately learns about an object category and a pose in an object-centric reference frame. We called our modular building block a Cortical Column Network, referring to the cortical column structures in the neocortex which (Hawkins et al., 2017) hypothesize also model objects in a local reference frame. However, despite the similarities, it is important to note that there are also important differences with how biological cortical columns are supposed to work in the Thousand Brains theory. For instance, each cortical column in the neocortex processes a distinct, small sensory patch, whereas our CCNs all work on the same, full resolution camera input. Moreover, each cortical column is hypothesized to model and vote for a larger number of object categories, which yields a more scalable processing architecture and sparse object representations. Finally, we also note that as the "what" and "where" information stream are located at distinct areas in the brain, this information is also processed through separate cortical columns. These and other aspects are not treated in our current CCN architecture, and it remains an exciting research direction to investigate to what extent artificial agents should mimick biologically plausible architectures and processing methods. For example, the representation of multiple internal models, or hypothesis for the sources of sensory information, has been explored in the context of birdsong and social exchange in the auditory domain (Isomura et al., 2019). Again, the basic idea is that multiple hypotheses are entertained and the model with the highest evidence contributes more to the posterior beliefs (i.e., Bayesian model average) about latent states or codes. Embedding our CCNs within the active inference framework enabled us to integrate both model learning and action selection under a single optimization objective. It would be interesting to investigate to what extent biological cortical columns could also be modeled to engage in active inference to produce motor commands.

Having a repertoire of object-specific cortical columns, who can "vote" or "compete" to explain sensory input, can be understood from a number of perspectives. The thousand brains perspective is closely related to mixtures of experts, of the kind found in MOSAIC architectures for motor control

(Haruno et al., 2003). Perhaps most generally, it can be regarded as a simple form of Bayesian model averaging (Hoeting et al., 1999). In other words, each cortical column builds a posterior belief about the attributes of an object, under the hypothesis or model that the object belongs to a particular class. The evidence for this model is then used to form a Bayesian model average over object attributes. A Bayesian model average simply marginalizes over the models by taking a mixture of posteriors under each model, weighted by the evidence for the respective models (usually, a softmax function of variational free energy).

Visual information is processed according to two major fiber bundles in the brain. It is hypothesized that these two fibers process separate aspects of the observed visual stimulus. The observed object identity is processed through a ventral pathway, and the objects location is processed through a specialized dorsal pathway (Grobstein, 1983). We found that using a ensemble of CCNs, a high accuracy classifier can be built combining both a ventral ("what") stream to infer object identity, and a dorsal ("where") stream inferring an object pose, and predicting other viewpoints. Crucially, we showed that an enactive, embodiment agent is required both to train such a system unsupervised, by collecting a dataset of viewpoints of each object, as well as to make correct inferences, and resolve ambiguity in the observation. In this work, we decoupled the data collection and the inference phase, and trained on a dataset of a relatively large number of randomly queried poses. Under active inference, one can also model the inference process over model parameters (Friston et al., 2016), and actively sample views that one expect to maximize information gain for the training process. It is worth investigating whether the model can be trained more efficiently, by driving the agent to the most informative view using the information gain on model parameters in the expected free energy functional. Information gain on model parameters is, in the active inference literature, called "novelty", while information gain on latent states or attributes is associated with "salience" (Schwartenbeck et al., 2019).

We also investigated the pose estimation properties using the dorsal ("where")-stream of our model. We showed that we are able drive the agent's actions toward a preferred, target pose by providing the corresponding observation. While we showed that the agent is indeed able to find a viewpoint with a similar observation, we also found that a lot of alias viewpoints exist in the latent space, due to object symmetries on the one hand, and the lack of sufficient visual details captured by the model, i.e., to disambiguate the front or back label of an object. However, we argue that in the case of robotic manipulation, this level of performance would already be sufficient for basic manipulation tasks such as grasping or pushing. Nevertheless it remains an important area of future work to find models that are able to capture and encode the required level of (visual) detail.

In addition there are still a couple of limitations in our current setup that might be addressed in future work. For example, our models currently learn the representation for a single object instance of an object category. In simulation, there is no variation between multiple instances of the mustard bottle, however, in real life the label can be attached crooked, or some markings can be present on the object. The current CCNs do not generalize to perturbations of the objects let alone other objects belonging to the same category, i.e., a coffee mug with a different height or color. It is worth investigating whether a single CCN can contain representations of different instances of a more general object category. Also note that our current CCN models are trained from scratch for each novel object category. Hence, a lot of overlapping information has to be relearned. Learning to re-use information would yield CCNs that are closer to the thousand brains theory as the cortical columns in the brain also reuse information (Hawkins et al., 2019). In order to re-use information learned by the CCNs, a potential extension would be to share weights between all CCNs for part of the layers, or devise a more hierarchical approach modeling part-whole relationships (Hinton, 2021).

Finally, our CCN models only encode egocentric representations in an object-local reference frame. In order to model a whole scene or workspace, the agent will need to map these egocentric poses into an allocentric reference frame (Parr et al., 2021). This would enable the agent to build a cognitive map of the workspace, inferring for each object an allocentric pose in the workspace, and "navigate" from one object to another. This would then give rise to a hierarchical generative model, mapping the world and its constituent objects using the same priniciples as simultaneuous localization and mapping (Safron et al., 2021).

## Related Work

In previous work, we built an artificial agent that learned such a generative model from pixel data, inferring beliefs about a latent variable representing the scene $\mathbf{s}$, given image observations $\mathbf{o_t}$ from absolute viewpoints $\mathbf{v}_t$ (Van de Maele et al., 2021b). Similar to a Generative Query Network (GQN) architecture (Eslami et al., 2018), this approach requires a huge train set of different scenes, with a limited set of constituent objects, in order to learn valid scene representations. The representations from this model encode all present objects and their relative pose with respect to the global allocentric reference frame. As a result, this lacks a factorization of different objects, and does not scale to a large number of objects present in the scene.

Most object-centric representations stem from the seminal work Attend Infer Repeat (AIR) by Eslami et al. (2016), where an image of a scene is factorized as a collection of latent variables separately describing the what and where parameters of each object. These variables are recurrently predicted, and can thus be scaled to an arbitrary amount of objects in the scene. AIR considers a static observer looking at a single observation. Burgess et al. (2019), proposed MONet, which learns the decomposition in an unsupervised end-to-end fashion. They also learn a structured representation describing each object. IODINE (Greff et al., 2019) also learns a joint decomposition and representation model but requires a fixed amount of slots that can be filled in by separate objects. Other work focuses on dynamic scenes by adding a temporal component (Kosiorek et al., 2018). They do this by adding a propagation module for objects from previous timesteps, and a discovery module that detects novel aspects. Other follow-up works tackle the scalability problem (Crawford and Pineau, 2020; Jiang et al., 2020) by predicting segmentation masks directly. Lin et al. (2020) combine

the scalability and temporal works, and add multimodality in the model through sampling in multiple steps. More recent work also considers three dimensional scenarios (Chen et al., 2021) with primitive shapes such as cubes or spheres where a generative query network (Eslami et al., 2018) is used as a rendering module for objects separately.

Similar to these models, we also make the separation in a what and where latent code. However, instead of forcing the factorization in a single latent space, we factorize on a model level, which results in a modular CCN model, where each CCN can focus on a single object type. While all mentioned models acquire impressive results on either static or dynamic observation data, none of these models consider an embodied, enactive agent to improve perception, which we believe to be crucial for intelligence.

An upcoming type of models are the implicit representation models that learn three dimensional structure explicitly in the model weights (Mescheder et al., 2019; Park et al., 2019; Mildenhall et al., 2020; Sitzmann et al., 2020). Neural Radiance Fields (NeRF) (Mildenhall et al., 2020), can learn complex object geometry by directly optimizing color and opacity values when conditioned by the coordinate and orientation of a point in the three dimensional space. This is optimized end-to-end directly from observation by casting rays from the camera pose and inferring sampled points on this ray. In follow-up work, different ways to optimize these models real-time by selecting key observations and strategic sampling of rays were found (Sucar et al., 2021). Similar to the implicit representation models, we learn to reconstruct object observations from a different set of observations. While the reconstruction detail of these models is impressive, these models lack an inverse model to infer poses or object categories.

A popular brain-inspired paradigm for unsupervised representation learning is predictive coding (Rao and Ballard, 1999). This mechanism works by hierarchically estimating the input and only propagating the error. This way, the lower levels of the hierarchy focus on smaller details of the observation. This work has also been used to separate the "what" and "where"-information streams (Rao and Ruderman, 1999). The predictive coding paradigm can be recast as active inference when using distributions over the predictions, rather than point estimates and when actions can be inferred to lead the artificial agent to a preferred goal state (Jiang and Rao, 2021).

The proposed approach in this article is also closely related to the object pose estimation research domain. These methods typically try to estimate the object pose directly as a 6 dimensional vector representing both the translation and orientation with respect to an absolute reference frame. Within the taxonomy provided in the survey paper by Du et al. (2021), our method could best be classified under the template-based label: given an observation, the model tries to find the pose that best matches one of the pre-defined labels. In this case, a trained CCN amortizes the process of finding the exact template through encoding the observation. The most closely related approaches use convolutional neural networks to directly estimate the object pose, and are pretrained on a set of labeled data which can be considered the templates (Do et al., 2018; Xiang et al., 2018; Liu et al., 2019). While these approaches acquire high accuracy results, they are trained supervised with a labeled dataset. In contrast, our approach is trained unsupervised from sequences of observations an enactive agent could perform, enabling our model to learn in arbitrary environments. This also has the corollary that the learned pose is in a non-interpretable latent space and can not be decomposed in an explicit translation and orientation.

The active inference (Friston et al., 2016) framework has also been previously adopted for describing generative models for active vision (Parr et al., 2021). In prior work, this framework has been shown to drive intelligent agents for visual foraging (Mirza et al., 2016; Heins et al., 2020), or for creating fovea-based attention maps to improve perception accuracy (Dauc, 2018). However, these works typically work with simpler, human engineered generative state space model, whereas in our case, the models are learned end-to-end from pixels. Different works also combine active inference with deep learning for learning state spaces directly using pixel-based observations (Çatal et al., 2020b; Fountas et al., 2020; Mazzaglia et al., 2021), but focused more on pixel-based benchmarks for reinforcement learning.

## 5. CONCLUSION

In this article, we proposed a novel method for learning object-level representations, drawing inspiration from the functional properties of the dorsal and ventral stream in the human neocortex. We made a separation on an object level, and create a basic building block for learning representations, which we coin a Cortical Column Network or CCN. We first described a generative model that casts vision as making inferences about an object pose and identity. For this generative model, we derived the (expected) free energy functional, which is used for both optimizing the model parameters as well as driving the agent actions toward desired poses or gaining information for better inference.

We showed that an ensemble of CCNs can be used for accurate object classification. By aggregating CCN predictions as "votes" in a Dirichlet distribution, we are able to correctly identify all known objects, while at the same time also being able to detect never seen before objects as an "other" category. We showed how an enactive, embodied agent improves the classification accuracy over time, by actively sampling novel observations that reduce ambiguity. We also investigated the ability of a CCN for reaching a preferred pose, given a target observation. We qualitatively evaluated how indeed the agent moves toward a matching observation. In addition, we explored the expected free energy landscape, showing that our models learn an abstract latent space for encoding pose in an object-local reference frame, exploiting object symmetries.

We believe that developing algorithms for learning in enactive, embodied agents is key to build artificial intelligent agents. To do so, we should rather inspire ourselves by the domains that study such embodied agents, i.e., behavioral psychology, biology and neuroscience, rather than only limit ourselves to the domain of artificial intelligence. We hope this work takes a small step in that direction.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## AUTHOR CONTRIBUTIONS

TVa and TVe conceived and performed the experiments and worked out the mathematical basis for the experiments. TVa, TVe, OÇ, and BD contributed to the manuscript. BD supervised the experiments. All authors contributed to the article and approved the submitted version.

## FUNDING

This research received funding from the Flemish Government (AI Research Program). OÇ was funded by a Ph.D. grant of the Flanders Research Foundation (FWO). Part of this work has been supported by Flanders Innovation & Entrepreneurship, by way of Grant Agreement HBC.2020.2347.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2022.840658/full#supplementary-material

## REFERENCES

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., et al. (2019). MONet: unsupervised scene decomposition and representation. *arXiv [Preprint] arXiv*:1901.11390. doi: 10.48550/arXiv.1901.11390

Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., and Dollar, A. M. (2015). "The ycb object and model set: towards common benchmarks for manipulation research," in *International Conference on Advanced Robotics (ICAR)*. (Istanbul), 510–517.

Çatal, O., Wauthier, S., De Boom, C., Verbelen, T., and Dhoedt, B. (2020a). Learning generative state space models for active inference. *Front. Comput. Neurosci.* 14, 574372. doi: 10.3389/fncom.2020.574372

Çatal, O., Wauthier, S., De Boom, C., Verbelen, T., and Dhoedt, B. (2020b). Learning generative state space models for active inference. *Front. Comput. Neurosci.* 14, 574372.

Chen, C., Deng, F., and Ahn, S. (2021). ROOTS: object-centric representation and rendering of 3D scenes. *arXiv [Preprint] arXiv*:2006.06130. Available online at: http://jmlr.org/papers/v22/20-1176.html

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., and Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosoph. Trans. R. Soc. B Biol. Sci.* 372, 20160055. doi: 10.1098/rstb.2016.0055

Crawford, E., and Pineau, J. (2020). "Exploiting spatial invariance for scalable unsupervised object tracking," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence* (New York, NY: AAAI Press), 3684–3692. Available online at: https://aaai.org/ojs/index.php/AAAI/article/view/5777

Dauc, E. (2018). Active fovea-based vision through computationally-effective model-based prediction. *Front. Neurorobot.* 12, 76. doi: 10.3389/fnbot.2018.00076

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural Comput.* 7, 889–904. doi: 10.1162/neco.1995.7.5.889

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (IEEE)* (Miami, FL), 248–255.

Do, T., Cai, M., Pham, T., and Reid, I. D. (2018). Deep-6dpose: recovering 6d object pose from a single RGB image. CoRR abs/1802.10367.

Du, G., Wang, K., Lian, S., and Zhao, K. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artif. Intell. Rev.* 54, 1677–1734. doi: 10.1007/s10462-020-09888-5

Eslami, S. M. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Kavukcuoglu, K., et al. (2016). "Attend, infer, repeat: Fast scene understanding with generative models," in *Advances in Neural Information Processing Systems*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates). Available online at: https://proceedings.neurips.cc/paper/2016/file/52947e0ade57a09e4a1386d08f17b656-Paper.pdf

Eslami, S. M. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., et al. (2018). Neural scene representation and rendering. *Science* 360, 1204–1210. doi: 10.1126/science.aar6170

Fountas, Z., Sajid, N., Mediano, P., and Friston, K. (2020). "Deep active inference agents using monte-carlo methods," in *Advances in Neural Information Processing Systems*, vol. 33, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc.), 11662–11675.

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2016). Wordbank: an open repository for developmental vocabulary data. *J. Child Lang.* 44, 677–694. doi: 10.1017/s0305000916000209

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Gilmer, J., Adams, R. P., Goodfellow, I. J., Andersen, D. G., and Dahl, G. E. (2018). Motivating the rules of the game for adversarial example research. CoRR abs/1807.06732.

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoren, D., et al. (2019). "Multi-object representation learning with iterative variational inference," in *Proceedings of the 36th International Conference on Machine Learning*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, CA: PMLR), 2424–2433. Available online at: http://proceedings.mlr.press/v97/greff19a.html

Grobstein, P. (1983). Analysis of visual behavior, david j. ingle, melvyn a. goodale, richard j.w. mansfield (eds.). mit press, cambridge, ma and london (1982), 834. *Anim. Behav.* 31, 621–622.

Ha, D., and Schmidhuber, J. (2018). "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems 31* (Montreal: Curran Associates, Inc.), 2451–2463.

Hanson, A. (1978). *Computer Vision Systems*. St. Louis, MI: Academic Press.

Haruno, M., Wolpert, D. M., and Kawato, M. (2003). Hierarchical MOSAIC for movement generation. *Int. Congr. Series* 1250, 575–590. doi: 10.1016/s0531-5131(03)00190-0

Hawkins, J., Ahmad, S., and Cui, Y. (2017). A theory of how columns in the neocortex enable learning the structure of the world. *Front. Neural Circuits* 11, 81. doi: 10.3389/fncir.2017.00081

Hawkins, J., Lewis, M., Klukas, M., Purdy, S., and Ahmad, S. (2019). A framework for intelligence and cortical function based on grid cells in the neocortex. *Front. Neural Circuits.* 12, 121. doi: 10.3389/fncir.2018.00121

Heins, R. C., Mirza, M. B., Parr, T., Friston, K., Kagan, I., and Pooresmaeili, A. (2020). Deep active inference and scene construction. *Front. Artif. Intell.* 3, 509354. doi: 10.3389/frai.2020.509354

Hinton, G. E. (2021). How to represent part-whole hierarchies in a neural network. CoRR abs/2102.12627

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Stat. Sci.* 14, 382–417. doi: 10.1214/ss/1009212519

Isomura, T., Parr, T., and Friston, K. (2019). Bayesian filtering with multiple internal models: toward a theory of social intelligence. *Neural Comput.* 31, 2390–2431. doi: 10.1162/neco_a_01239

James, K. H., Jones, S. S., Smith, L. B., and Swain, S. N. (2014). Young children's self-generated object views and object recognition. *J. Cogn. Develop.* 15, 393–401. doi: 10.1080/15248372.2012.749481

Janssen, P., Vogels, R., and Orban, G. A. (2000). Three-dimensional shape coding in inferior temporal cortex. *Neuron.* 27, 385–397. doi: 10.1016/s0896-6273(00)00045-3

Jiang, J., Janghorbani, S., Melo, G., and Ahn, S. (2020). "SCALOR: Generative world models with scalable object representations," in *8th International Conference on Learning Representations* (Addis Ababa: OpenReview). Available online at: https://openreview.net/forum?id=SJxrKgStDH

Jiang, L. P., and Rao, R. P. N. (2021). Predictive coding theories of cortical function. *arXiv [Preprint]*. arXiv: 2112.10048. Available online at: https://arxiv.org/ftp/arxiv/papers/2112/2112.10048.pdf

Kingma, D. P., and Ba, J. (2015). "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, eds Y. Bengio and Y. LeCun (San Diego, CA). Available online at: http://arxiv.org/abs/1412.6980

Kingma, D. P., and Welling, M. (2014). "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations*, eds Y. Bengio and Y. LeCun (Banff, AB). Available online at: http://arxiv.org/abs/1312.6114

Kosiorek, A. R., Kim, H., Teh, Y. W., and Posner, I. (2018). "Sequential attend, infer, repeat: Generative modelling of moving objects," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, eds S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: NeurIPS), 8615–8625. Available online at: https://proceedings.neurips.cc/paper/2018/hash/7417744a2bac776fabe5a09b21c707a2-Abstract.html

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. in *Proceedings of the 25th International Conference on Neural Information Processing 757 Systems - Volume 1, NIPS'12* (Red Hook, NY: Curran Associates Inc.), 1097–1105.

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cogn. Develop.* 3, 299–321. doi: 10.1016/0885-2014(88)90014-7

Lin, Z., Wu, Y. -F., Peri, S. V., Fu, B., Jiang, J., and Ahn, S. (2020). "Improving generative imagination in object-centric world models," in *Proceedings of the 37th International Conference on Machine Learning (PMLR)*, 6140–6149. Available online at: http://proceedings.mlr.press/v119/lin20f.html

Liu, F., Fang, P., Yao, Z., and Yang, H. (2019). Recovering 6 d object pose from rgb indoor image based on two-stage detection network with multitask loss.

Mahajan, D., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., et al. (2018). Exploring the limits of weakly supervised pretraining. CoRR abs/1805.00932.

Mazzaglia, P., Verbelen, T., and Dhoedt, B. (2021). "Contrastive active inference," in *Advances in Neural Information Processing Systems (Online)*.

Mescheder, L. M., Oechsle, M., Niemeyer, M., Nowozin, S., and Deiger, A. (2019). "Occupancy networks: Learning 3D reconstruction in function space," in *IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 4460–4470. doi: 10.1109/CVPR.2019.00459

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoothi, R., and Ng, R. (2020). NeRF: representing scenes as neural radiance fields for view synthesis. *Commun*. ACM 65, 99–106. doi: 10.1145/3503250

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE* 13, e0190429. doi: 10.1371/journal.pone

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci.* 10. doi: 10.3389/fncom.2016.00056

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* 6, 414–417. doi: 10.1016/0166-2236(83)90190-x

Mountcastle, V. (1997). The columnar organization of the neocortex. *Brain* 120, 701–722. doi: 10.1093/brain/120.4.701

Park, J. J., Florence, P., Straub, J., Newcombe, R. A., and Lovegrove, S. (2019). "DeepSDF: Learning continuous signed distance functions for shape representation," in IEEE Conference on Computer Vision and Pattern Recognition (Long Beach, CA: IEEE), 165–174. doi: 10.1109/CVPR.2019.00025

Parr, T. and Friston, K. J. (2018). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12. doi: 10.3389/fncom.2018.00090

Parr, T., Sajid, N., Da Costa, L., Mirza, M. B., and Friston, K. J. (2021). Generative models for active vision. *Front. Neurorobot.* 15, 651432. doi: 10.3389/fnbot.2021.651432

Rao, R., and Ruderman, D. (1999). "Learning lie groups for invariant visual perception," in *Advances in Neural Information Processing Systems*, Vol. 11, eds M. Kearns, S. Solla, and D. Cohn (Denver, CO: MIT Press).

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31th International Conference on Machine Learning* (Beijing: JMLR), 1278–1286. Available online at: http://proceedings.mlr.press/v32/rezende14.html

Rezende, D. J., and Viola, F. (2018). Taming VAEs. *arXiv [Preprint] arXiv*:1810.00597. doi: 10.48550/arXiv.1810.00597

Safron, A. (2021). The radically embodied conscious cybernetic bayesian brain: from free energy to free will and back again. *Entropy*. 23. doi: 10.3390/e23060783

Safron, A., Çatal, O., and Verbelen, T. (2021). Generalized simultaneous localization and mapping (G-SLAM) as unification framework for natural and artificial intelligences: towards reverse engineering the hippocampal/entorhinal system and principles of high-level cognition. *PsyArXiv. Preprints*. doi: 10.31234/osf.io/tdw82

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., and Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife.* 8, e41703. doi: 10.7554/eLife.41703

Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., and Wetzstein, G. (2020). "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. -F. Balcan, and H. -T. Lin. Available online at: https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html

Smith, L. B., Yu, C., and Pereira, A. F. (2010). Not your mother's view: the dynamics of toddler visual experience. *Develop. Sci.* 14, 9–17. doi: 10.1111/j.1467-7687.2009.00947.x

Smith, R., Friston, K. J., and Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *J. Math. Psychol.* 107, 102632. doi: 10.1016/j.jmp.2021.102632

Sucar, E., Liu, S., Ortiz, J., and Davison, A. (2021). "iMAP: Implicit mapping and positioning in real-time," in *Proceedings of the International Conference on Computer Vision*.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. (2021). Deep learning's diminishing returns: the cost of improvement is becoming unsustainable. *IEEE Spectr.* 58, 50–55. doi: 10.1109/MSPEC

Tse, P. U. (1999). Volume completion. *Cogn. Psychol.* 39, 37–68. doi: 10.1006/cogp.1999.0715

Van de Maele, T., Verbelen, T., Çatal, O., De Boom, C., and Dhoedt, B. (2021b). Active vision for robot manipulators using the free energy principle. *Front. Neurorobot.* 15, 642780. doi: 10.3389/fnbot.2021.642780

Van de Maele, T., Verbelen, T., Çatal, O., and Dhoedt, B. (2021a). "Disentangling what and where for 3d object-centric representations through active inference," in *2nd International Workshop on Active Inference (IWAI)*.

von Helmholtz, H. (1977). "The facts in perception," in *Boston Studies in the Philosophy of Science.* (Netherlands: Springer), 115–185.

Xiang, Y., Schmidt, T., Narayanan, V., and Fox, D. (2018). "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Robotics: Science and Systems XIV, Carnegie Mellon University*, eds H. Kress-Gazit, S. S. Srinivasa, T. Howard, and N. Atanasov (Pittsburgh, PA). doi: 10.15607/RSS.2018.XIV.019

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2012). "A comprehensive evaluation of full reference image quality assessment algorithms," in *19th IEEE International Conference on Image Processing*. (Orlando, FL), 1477–1480.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Van de Maele, Verbelen, Çatal and Dhoedt. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

REVIEW
published: 25 April 2022
doi: 10.3389/fnbot.2022.857614

# Self-Concern Across Scales: A Biologically Inspired Direction for Embodied Artificial Intelligence

*Matthew Sims**

*Institute for Philosophy II, Ruhr-University Bochum, Bochum, Germany*

Intelligence in current AI research is measured according to designer-assigned tasks that lack any relevance for an agent itself. As such, tasks and their evaluation reveal a lot more about our intelligence than the possible intelligence of agents that we design and evaluate. As a possible first step in remedying this, this article introduces the notion of "self-concern," a property of a complex system that describes its tendency to bring about states that are compatible with its continued self-maintenance. Self-concern, as argued, is the foundation of the kind of basic intelligence found across all biological systems, because it reflects any such system's existential task of continued viability. This article aims to cautiously progress a few steps closer to a better understanding of some necessary organisational conditions that are central to self-concern in biological systems. By emulating these conditions in embodied AI, perhaps something like genuine self-concern can be implemented in machines, bringing AI one step closer to its original goal of emulating human-like intelligence.

**Keywords: homeostasis, embodied cognition, anticipatory control, artificial intelligence, artificial symbioses, basal cognition, common fate, goal directed behaviour**

OPEN ACCESS

***Edited by:***
*Adam Safron,*
*Johns Hopkins Medicine,*
*United States*

***Reviewed by:***
*Tim Verbelen,*
*Ghent University, Belgium*
*Mark Daniel Miller,*
*Hokkaido University, Japan*

***Correspondence:**
*Matthew Sims*
*matthew.sims-m4e@RUB.de*

***Received:*** *18 January 2022*
***Accepted:*** *16 March 2022*
***Published:*** *25 April 2022*

***Citation:***
*Sims M (2022) Self-Concern Across Scales: A Biologically Inspired Direction for Embodied Artificial Intelligence. Front. Neurorobot. 16:857614. doi: 10.3389/fnbot.2022.857614*

## INTRODUCTION

Artificial intelligence (AI) was originally described as the project of making a machine behave in ways that would be called intelligent if a human were so behaving (McCarthy et al., 1955). Central to this notion of intelligence is the idea of task evaluation. True intelligent behaviour is read off from an agent's ability to successfully complete tasks requiring something akin to human cognitive capacities to be successfully completed.[1] AI has generally fallen into two categories that align with two classes of tasks. The first category is *specialised AI*, which is designed with the aim of carrying out and being evaluated with respect to very *specific tasks* (e.g., playing GO, driving cars, generating language, etc.). *General AI*, on the other hand, is designed to carry out a broad domain of tasks that, at the time of design, are largely unknown (Thórisson et al., 2016). Although there is no agreed upon notion of intelligence in the AI literature, it is task evaluation, something that is often based on human psychological metrics that are used to determine whether a performance qualifies as intelligent or not. The Turing Test is a more general illustrative example of this manner of framing the concept of intelligence around the completion of tasks that are evaluated using human-based psychometrics (i.e., an artificial agent is intelligent if it can respond to a series of questions in a manner that is indistinguishable from responses of a human agent) (Turing, 1950).

[1] Although "intelligence" is often used to refer to higher cognitive capacities in the literature, in what follows, the terms "intelligence" and "cognition" will be used synonymously.

Regardless of whether an AI agent has been designed to complete specialised tasks or even in the case of general AI, where tasks might not be fully known when designing the system, the general domain of tasks must eventually be recognised by designers to be evaluated. This suggests that tasks and their evaluation reveal a lot more about our intelligence than the possible intelligence of agents that we design and evaluate. Man and Damasio (2019), thus, ask the following important question regarding the legitimacy of using the completion of designer-based tasks as a reliable indicator of human-like intelligence:

> "*Whose goals*?" Does an agent that myopically follows orders to the extent that it endangers itself and compromises its ability to carry out future orders deserve to be called intelligent?" (p. 447; author's emphasis).

An agent, for instance, that is equipped with a deep convolutional network and is able to correctly classify instances of threatening dogs when they are present 99% of the time in various environments but that fails in the capacity to behave in ways that allow it to avoid being damaged by a threatening dog also seems to lack something central to basic intelligence. Although such an agent might continue to complete the "visual" classification task that it has been given, the fact the completion of that task is compatible with the agent blindly pursuing its own destruction is at odds with the intuitive idea that basic intelligent behaviour is something that in living systems is typically adaptive and supports self-maintenance.

There is another approach to tasks and intelligence, however. This alternative approach suggests evaluating intelligence according to how a system completes tasks that bring about its own goals. It is becoming increasingly recognised that plants (Baluška and Mancuso, 2009; Shemesh et al., 2010; Trewavas, 2014; Gagliano et al., 2016; Novoplansky, 2016; Calvo et al., 2020) and basal systems (Maturana and Varela, 1980; Nakagaki et al., 2000; Hellingwerf, 2005; Ben Jacob et al., 2006; Lyon, 2006; Van Duijn et al., 2006; Shapiro, 2007; Saigusa et al., 2008; Baluška and Levin, 2016; Pinto and Mascher, 2016; Reid and Latty, 2016; Levin, 2019; Bechtel and Bich, 2021; Boussard et al., 2021; Hanson, 2021; Lyon et al., 2021) display some degree of intelligence that is expressed in various manners in which they adapt to the complexity of their environments. Selective pressures and environmental stresses that challenge both homeostasis and development are fundamental existential tasks that all biological systems encounter and must adaptively respond to. Importantly, remaining in a limited and select range of viability supporting physiological states is something that every biological system has a concern for (Jonas, 1966/2001). As such, any biological system is motivated from its genesis to deal with existential tasks that are intrinsic to it (Barrett, 2019). A system's own tasks are an expression of what Kant called "purposiveness without purpose" (1790/2007).[2] All other tasks are organised around the task of continued self-maintenance, a task that endows both the environment and an organism's environmental interactions with meaning from its perspective.

The tendency of a system to bring about states that are compatible with its continued self-maintenance when perturbed is what I shall refer to as self-concern. Because self-concern is underwritten by a system's ability to measure and track the evolution of its own physical states and compare these with encoded optimal states, such concern may be said to reflect a form of self-reference. That said, self-concern should not be understood as involving the occasioning of personal-level states (e.g., beliefs and/or desires of folk-psychology) about the self or self-awareness. As it is being envisioned here, self-concern is strictly a subpersonal-level phenomenon, which can diverge from concern at the personal level (e.g., various systems that one's body is composed may slowly fail to track and return to states that are compatible with one's continued viability due to senescence; however, one during this period of development may be concerned for one's own survival). Importantly, if it is assumed that this kind of concern forms the basis of biological intelligence (something that will later be argued is, in fact, a reasonable assumption) and that biological intelligence offers a powerful and revealing lens with which to view intelligence across the board, then the following question becomes an emphatic one: how do we go about designing self-concerned AI agents in a way that sets the stage for basic intelligence to emerge in such agents?

The aim of this article is to investigate and bring to light some features of the relationship between intelligence and self-concern in biological systems in a manner that can be used to inform AI research. Self-concern in biological intelligent systems, it shall be argued, presupposes some form of embodiment. It is largely in part due to an environment's long-term effects (beneficial or adverse) on a system's body encountered through the interface of that body that its environment comes to have a meaning for that system, a meaning that is reflected in viability-sustaining behaviour driven by self-concern. Furthermore, various intelligent adaptive strategies that contribute to continued metabolic functioning and boundary regulation, it shall be argued, are tantamount to a series of solutions to a system's own intrinsic task (i.e., goal) of continued self-maintenance. It will be argued that until agents are able to exhibit something like self-concern on multiple scales of their embodiment, the kind of specialised and general tasks used to evaluate AI will continue to be overly theory-laden reflections of our own intelligence. Otherwise stated, until AI agents are concerned about their continued self-maintenance, their behaviour will continue to be exclusively guided by the processing of syntactical information that has meaning to us rather than semantic information that is grounded in an AI agent's own viability conditions (Kolchinsky and Wolpert, 2018). My aim is to cautiously progress a few steps closer to a better understanding of some organisational conditions that are central to self-concern in biological systems. By emulating these conditions (and most likely some others) in embodied AI, perhaps something like genuine self-concern in machines can emerge, bringing AI one step closer to its original goal.

[2]This is the idea that the worth of a purposive system resides in its own being rather than a result, the value of which is determined external to that system (Cassirer, 1981).

The rest of the article is organised as follows: Section A Biologically Inspired Functional Approach to Intelligence provides a description of the biologically inspired functional approach to intelligence and then turns to the central notion of concern, showing how intelligence and concern are related; Section Three Necessary Features for Self-Concern in Biological Systems focuses on three necessary features of concern found in living systems: "system-environment energetic traffic," the system-initiated process of energy accrual and exchange that allows an agent to defend its systemic boundaries and maintain its functional organisation *via* self-production; "dual information-carrying nature of interfacing bodily elements," the property of an embodied system to harvest information simultaneously about the states of its external environment and its internal states; "hierarchically structured systems that share a degree of common fate," the spatially and temporally nested organisation of systemic parts, the interaction of which is mutually supportive on different timescales; Section When Anticipatory Dynamics Answer to Self-Concern then looks at how anticipatory behaviour is related to concern and what this means for designing embodied AI systems in a manner that will get them closer to being self-concerned adaptive systems. I conclude with some brief remarks about some various design challenges and ethical issues that arise when considering self-concern in AI.

## A BIOLOGICALLY INSPIRED FUNCTIONAL APPROACH TO INTELLIGENCE

One manner of restricting the scope of tasks (and task evaluation) in a way that respects the fact that intelligent behaviour is, at its core, an adaptive strategy is to deploy a biological functional approach to the notion of intelligence when designing AI. Such a biologically inspired approach describes what intelligence does for an agent as opposed to defining intelligence relative to the purposes and interests of the designer. For example, Sejnowski (2018), in describing intelligence, states that it is a general capacity that "evolved in many species to solve the problems they faced to survive in their environmental niches" (p. 263). Intelligence, from this perspective, is seen, first and foremost, as a solution to a (moving) set of environmental challenges that a system must adaptively respond to in order to remain alive[3]. To be sure, this functional approach does not claim that intelligence is limited to solving survival-related environmental problems; rather, it makes a more modest claim that intelligence is, fundamentally, a strategy for coping with environmental complexity (Godfrey-Smith, 1996; cf. Lyon, 2006).

Importantly, characterising intelligence in terms of its evolutionary function means recognising that it is something that can only be defined in relation to the kind of environment that an embodied agent must deal with in order to survive[4]. Being embodied and embedded in an environment is a precondition for cognition and anything that might be accurately deemed intelligent behaviour (Bateson, 1972; Clark, 1997; Pfeifer and Bongard, 2006; Pfeifer et al., 2007; Pezzulo et al., 2011; Lara et al., 2018). As such, differences in agent morphology and niche will be reflected in different forms of intelligent behaviour. For example, human intelligence, given the specifics of the human niche, will be different in form than, say, the intelligence that allows bees to successfully navigate their bee niches, pursuing opportunities for action, what Gibson (1966) called "affordances," and avoiding potential harm-inducing situations that are specific to bee-like animals. In characterising biological intelligence relative to an agent's continued survival in its environmental niche, the notion of intelligence is rendered a relational property as opposed to a capacity that can be understood (or investigated) in the abstraction of what intelligence is a response to. In approaching intelligence by way of its biological function, we glean some insight (with a fair amount of speculation of course) into conditions under which various forms of intelligence have been evolutionarily selected for and, hence, why such forms of intelligence (human or non-human) are present today.

Crucially, by deploying a biologically inspired functional approach to intelligence, AI is not restricted to using human intelligent behaviour as a gold standard; although the aim of AI may be to design agents whose intelligent behaviour is human-like, a functional approach to intelligence suggests a different (yet not exclusively so) starting point to the investigation of intelligence. In taking a "biogenic approach" (Lyon, 2006) and, thus, recognising cognition's fundamental role as an evolved adaptive strategy, one first carries out a detailed investigation of the simplest instances of intelligent behaviour in basal biological systems and then works up to the more complex cases of intelligence in humans. Intelligence is, thus, viewed as something that reflects an evolutionary continuity among different forms of life along common phylogenetic branches and across different branches; this is a recognition that various forms of cognition may be evolutionarily convergent strategies that have arisen many times in different phyla much like vision or breathing has arisen multiple times, taking different forms. Human intelligence is, thus, not different in kind from that of simple organisms but different in form (cf. Darwin, 1871). Such a functional approach to intelligence provides a much-needed conceptual and methodological basis for throwing light on the fact that there is a "spectrum of intelligent behaviour found in nature that artificial systems can learn from" (cf. Webb and Scutt, 2000; Dupeyroux et al., 2017; Sejnowski, 2018, p. 267).

One apparent snag in deploying a functional approach to inform the development of AI is this: AI agents, even when taking the form of embodied agents, are typically fully abiotic

[3]Locating the notion of human intelligence in various abilities to cope with survival-relevant human environmental tasks, the aim of AI, as it was introduced by McCarthy et al. (1955), might be reframed as the project of getting machines to cope with human niches in ways that would be recognized as being analogous to how humans adapt to the complexity of our niches.

[4]The famous cyberneticist Bateson (1972), in many ways, auguring the arrival of embodied and situated cognition, recognized the importance of considering bodies and environments as parts of cognitive explanation. According to him, the cognitive system is a unit of explanation, the bounds of which are determined by information-carrying pathways (i.e., circuits) that cannot be severed without rendering the explanandum mysterious.

agents. They are neither subject to heritable mutation nor does the notion of selective fitness apply to them; they are neither subject to senescence nor do suffer death. Of course, evolutionary algorithms might simulate such processes, but the chosen parameters that define the simulation and the chosen desiderata used to score fitness in such simulations are just that, chosen by designers (but see Bongard and Levin, 2021; Lehman and Stanley, 2011). One of the reasons that is embedded in an environment might be thought of as a precondition for intelligent behaviour is the fact that the presence of uncertainty that is inherent in natural environments has real existential consequences for systems situated in these environments. Fitness, which may be construed as an inverse cost function, "is a concept that only has meaning in the context of a concrete set of constraints, either from the environment or from the system being optimised" (Sejnowski, 2018, p. 267). Conceding to all of this is, however, consistent with holding the claim that basing the notion of intelligence on adaptation does not place any actual biological requirements on AI. It is not necessary that an agent undergoes or partakes in exact processes that are central to adaptation in living systems; rather, in order for machines to exhibit intelligence in any manner comparable to the simplest expressions of biological intelligence (e.g., archaea, bacteria, slime mould, yeasts, living cells, etc.), it may be the case that they will have to engage in some processes that are functionally similar to those that are characteristic of living systems. The presence of such biologically inspired processes, I shall argue, is the bedrock on which self-concern in machines might arise.[5]

## Self-Concern in Intelligent Systems

Returning to Man and Damasio's point above a central aspect of intelligent behaviour is that the goals of a behaving system can be said to be that system's own. Goals may be construed as monitorable physical states that a system's continued functioning depends on and to which a self-organising system tends to return after perturbation. In living systems, such goals may be construed as "homeostatic imperatives" (Man and Damasio, 2019), which a system's self-concern is defined with respect to; it is a system's goal of self-maintenance, which endows its actions and the environment with meaning from its perspective. Deploying concepts from cybernetic control theory (Wiener, 1948; Ashby, 1952; Conant and Ashby, 1970; Bateson, 1972), self-concern can be associated with a regulating system's tendency to return to an optimal set point range that is consistent with its continued functioning.[6] It is the concern that an agent has for stabilisation of its physiological states within its viable set point range that motivates not only the simplest expressions of intelligent behaviour but also acts as the foundation on which the most complex forms of intelligent behaviour are grounded.[7]

Behavioural avenues of self-concern mirror the open-ended definition of evolution. There are myriad ways that such concern might be behaviourally expressed, and the update of options available to mitigate self-concern at any given time is the result of a system's continuous dynamic exchange with its environment. To put it differently, self-concern is the driver of flexible and evolving behavioural solutions to homeostatic challenges posed by hostile environments. As such, it reflects the inherent creativity that some have acknowledged to be central to cognition and life (Kant, 1790/2007; Goodwin, 1978, 1994).

Although various architectures implementing theories such as optimal control theory (Berridge and Robinson, 2003; Sterling, 2012), drive reduction theory (Hull, 1943; Konidaris and Barto, 2006), and homeostatic reinforcement learning (Sutton and Barto, 1998; Oudeyer and Kaplan, 2007; Keramati and Gutkin, 2014) have placed homeostatic maintenance state front and centre as a driver of intelligent behaviour in agents, self-concern fails to be directly addressed in current AI research.[8] Part of this may be due to the fact that the relationship between self-concern and intelligent behaviour in biological systems is, itself, poorly understood. One of the aims of this article is to correct this; integral to the development of basic intelligence in AI, the kind that is ubiquitous in the living world, is understanding the details of self-concern in biological systems and designing agents that can engage in anticipatory homeostatic error correction fuelled by functionally similar machine self-concern. The next section offers a possible starting point for such a biologically inspired approach to embodied AI.

# THREE NECESSARY FEATURES FOR SELF-CONCERN IN BIOLOGICAL SYSTEMS

In what follows, I shall present three related features that I will argue are necessary (but not sufficient) for the emergence of self-concern in biological systems. These features are:

- Controlled system-environment energy traffic.
- Dual information-carrying nature of interfacing bodily elements.
- The common fate of hierarchically structured systems.

It is my hope that by making these features explicit, they may be instructive for the designing and improvement of already existing designs of biologically inspired agents, advancing the field of embodied AI one step closer towards the emergence of self-concern in artificial agents.

[5]This is particularly telling: if something like adaptivity is assumed to be a fundamental feature of all intelligent behaviours, then the progress of AI will not proceed the progress of artificial life (AL) or vice versa; the progress of both the AL and AI programmes will depend on their marching in lockstep.

[6]The notion of stable set points borrowed from the cybernetic control theory is a useful abstraction in the sense that homeostatic equilibria are not stable but shift across the lifetime of a biological system. Hence, homeostatic set points are more accurately seen as dynamic or moving equilibria.

[7]The fact that homeostasis is emphasized in this article as a basis for self-concern, however, does not imply that the process of heterostasis (i.e., exploring and seeking high information gain *via* self-perturbation) is not a significant driver of behaviour in self-concerned living systems.

[8]Oudeyer and Kaplan (2007) focus on the development of models of "intrinsic motivation" within the framework of reinforcement learning, and there may be some general overlap with what I am calling self-concern.

## Controlled System-Environment Energy Traffic

The first feature that I will argue is central to self-concern in biological systems, controlled system-environment energy traffic. The specific form that this traffic takes in biological systems occurs in the service of metabolism, which is "the set of life-sustaining chemical reactions within living cells" (Lane, 2016, p. 295). The two types of reaction characteristic of metabolism across all life forms are "anabolic reactions" (i.e., storing energy in the form of synthesised adenosine triphosphate, ATP) and "catabolic reactions" (i.e., breaking down of ATP into ADP + Pto release energy for work). It is through the acquisition of resources (e.g., nutrients) that living systems, *via* the combination of these reactions, are able to maintain the physiological processes which underwrite their ability to remain far from thermodynamic equilibrium with their environments, temporarily flouting the second law of thermodynamics. For biotic self-organising systems, remaining far from thermodynamic equilibrium means remaining alive (Nicolis and Prigogine, 1977; Friston, 2012, 2019; Demirel, 2014). A system, by harvesting resources from its local environment, can fuel the metabolic processes that allow it to (a) generate itself materially (e.g., protein synthesis) and (b) maintain its organisation in the face of environmental perturbation and despite continuous material turnover; these are the respective processes of "self-production" and "self-maintenance," which form the basis of Maturana and Varela's (1987) notion of "autopoiesis" (see also Gánti, 2003, Chemoton model).

It is the sense in which a system both actively pursues resources and actively directs how these resources are used (i.e., how self-production and self-maintenance play out) that system-environment energy traffic is controlled (cf. Bechtel and Bich, 2021). Why might the process of controlled energic traffic be required for self-concern? Phenomenologist Hans Jonas provides a hint when he writes:

> "In order to change matter, the living form must have matter at its disposal, and it finds it outside of itself, in the foreign 'world.' Thereby life is turned outward and towards the world in a peculiar relatedness of dependence and possibility. It wants to go out to where its means of satisfaction lie: its self-concern, active in the acquisition of new matter, is essential openness for the encounter of outer being." (Jonas, 1966/2001, p. 84).

The core idea that Jonas so eloquently expresses is that one of the directions of energy traffic that metabolism presupposes (taking in raw materials as a source of energy) places any living system in a relationship of need with its milieu; this need fundamentally arises from the fact that living systems are subject to constant material turnover that can only occur when that system acquires new resources. As such, this need implies a concern on the part of the behaving system for fulfilling its metabolic demands. Taking this into consideration, we may say that controlled system-environment energy traffic presupposes certain dependence on the environment on the part of the traffic controlling system. This dependence suggests that these systems that seek out sources of energy that they, in turn (*via* metabolism), use for work have a basic concern for their continued existence. Such is the primitive goal that every living system is concerned to satisfy, a goal that intelligent behaviour answers to and that provides the metabolising energy trafficking system with a basic perspective on the world (Lyon, 2006; Lyon et al., 2021). If this claim is in the right ballpark, then how may it be used to inform the development of AI?

From what has been said, we may glean this: if an agent altogether lacks the need for system-environment energy traffic that allows it to both maintain and produce itself to some degree, then that agent also lacks the capacity to exhibit basic concern. Such an agent fails to be autonomous (cf. Kauffman, 2000). Any behaviour that may be usefully ascribed to it, no matter how intelligent such an agent may be judged to be, fails to be its behaviour, because it does not stem from or, importantly, answer to its own concern. To be sure, both biological systems and machines require energy to do useful work in any capacity. However, and this difference is telling, biological systems both constantly monitor their energy levels across different spatial and temporal scales and actively behave in ways to fulfil their energetic needs. Such behaviour is not merely foraging for sources of usable energy but also generating the very materials and processes that allow for such energy foraging to continue to occur. Moreover, although to different degrees, each nested constituent part of a biological system, when all is going well, both contributes to and benefits from controlled energy traffic. This multi-scale concern reflects a basic "self-similarity" that is unique (at least for now) to the hierarchical organisation of biological systems (Bongard and Levin, 2021) (more will be said about this below). It is for this reason that an individual's biological parts (i.e., cell, organs, etc.), each of which takes part in its own energetic processing (i.e., monitoring and regulation), allow for distributed (decentralised) control at a very fundamental level[9].

This brings us to the following question: could a machine ever truly engage in controlled system-environment energy traffic with its environment? Sure, but probably not in the same manner that an organism can. Metabolism occurs on the nanoscale where "there is spontaneous motion, but there is enough structure and the relations between forces are such that a lot can happen, by biassing tendencies in random walks" (Godfrey-Smith, 2016, p. 5). The rapid development of nanobots (Berger, 2016; Service, 2016; Linke et al., 2020) and engineered nanomaterials (Galetti et al., 2019) suggests that operation on the nanoscale itself, however, fails to present an uncrossable boundary to

[9]One dimension in which system-environment energy traffic might vary in degree is with respect to how much each constituent part of a system both contributes to and relies on the energy traffic of the system (s) in which it is nested or coupled to. For example, if system-environment energy traffic is limited to only one nested element of a larger nesting system, then that larger system might satisfy the requirement of system-environment energy traffic but only to a minimal degree, whereas if each constituent and nested element of a larger system contributes to and relies on the energy traffic of all other elements (to varying degrees), the supraordinate system has a high degree of system-environment energy traffic. Quantifying this dimension of system-environment energy traffic falls out of the scope of this paper.

imitating energy traffic in machines (but see Nicholson, 2020).[10] Furthermore, ongoing advancements in self-replicating robots (for a detailed review, see Moses and Chirikjian, 2020) may be a promising starting point for the development of a basic form of artificial self-production.[11] It is a starting point because there is a notable difference between what might roughly be self-replication in machines and self-producing machines; whereas the former involves the use of supplied raw building materials to produce other machines that are copies of themselves, the latter involves the continuous generation (synthesis) of a system's own parts. To date, self-replicating robots are still limited to using supplied resources and cannot self-produce these materials from the ground up (Schranz et al., 2020); in other words, they lack the kind of "operational closure" that is unique to autonomous autopoietic systems (Maturana and Varela, 1980).[12]

One thing, however, that is key to keep in mind is that a system that is not subject to material degeneration (at least on short timescales) is a system that does not require (artificial) self-production. In other words, when an agent can both undergo substantial wear and tear due to its behaviour and environmental perturbations and monitor its own material degeneration, then harvesting energy from environmental resources takes on a particular value for the system; it is a manner of contributing to its own persistence across material turnover. Such turnover is something that is ultimately linked to the materiality of an embodied agent. For this reason, a soft robotic implementation may be invaluable in providing some implementational conditions for material turnover to occur and, hence, provide a need for artificial self-production. The production and development of microbial fuel cells (MFCs) in robots (Ieropoulos et al., 2005; Philamore et al., 2016) seem to be a promising manner of getting energetically autonomous embodied AI off the ground *via* biological metabolism (more on this below). MFCs provide systems that use them with conditions for material turnover and a need for self-production (at the level of microbes in the fuel cell).

Does it matter that "artificial metabolism" in machines will most likely be quite different from biological metabolism when it comes to being the fundamental of an agent's concern? I would like to suggest that such a question should not be addressed prior to the advent of artificial metabolism in agents. If it is telling of anything, to decide beforehand reveals nothing more than a deep commitment to the use of *a priori* intuitions, intuitions that may or may not be hostage to a deeply ingrained "biocentrism" when it comes to metabolism and/or self-concern (cf. Meincke, 2018).

Let us now turn to the second necessary feature of self-concern in biological systems, which, I will argue, is also a prerequisite for basic intelligence in embodied AI.

## Dual Information-Carrying Nature of Interfacing Bodily Elements

Even at the most basal level, biological self-concern relies on a system having multiple sources of feedback from its internal and external environments. It is only through the evaluation of such feedback that a system's behaviour may be directed in one way or another to return it to states that are compatible with its continued existence (i.e., its set-point values). Self-concern, although a feature of an entire system, is something that each component part of a biological system contributes to *via* the registration of information regarding its current state or condition and the state of the environment that it interacts with. Of particular importance are the states of components that causally interface the system with its external milieu.

From bacteria to humans, the presence of some form of the membrane that separates a living system from its external environment is ubiquitous. For example, *Escherichia coli* has a cell membrane that acts as a boundary between its cytoplasm and the external (terrestrial or fluid) medium, the dynamics of which the bacterium must behaviourally adapt to. The cell membrane and the flagellar motor machinery that it houses not only transmit external mechanical forces (tension, compression, and shear) to the internal components of the cell (Dufrêne and Persat, 2020), but multiple kinds of transmembrane receptor proteins and dedicated sensors also allow environmental feedback in the form of chemical gradients (Macnab and Koshland, 1972), light (Fraikina et al., 2015), and mechanical force (Dufrêne and Persat, 2020). Importantly, it is *via* the effects of environmental stimuli on the cell membrane and other environment interfacing components (e.g., flagella) that various internal biochemical cascades arise that contribute to the bacteria's ability to cope with their environmental dynamics. Were this interface to become non-responsive, the *E. coli*'s responses to its environment would in effect become random and ineffective, and thus the bacterium would soon cease to be. Like the simple *E. coli*, all living systems' membranes act as a conduit for proximal information about their environment. This proximal information is relevant to the notion of concern because its arising presupposes that some external force or stimulus has made a sensory contact with the organism already by way of inducing a change in the state of the membrane.

Of particular importance is the presence of mechano-stimulation, because it requires direct contact with the membrane or other bodily components that interface with the environment (e.g., hair, feathers, antennae, flagella, cilia, etc.). Proximal stimulus detection is not only informative about the states of the world but also the states of the system itself. If an abnormally high amount of mechanical stress is exerted upon the plasma membrane of a eukaryotic cell, it would result in membrane breach and that cell's likely death (Cooper and McNeil, 2015). Less dramatically, too much concentrated friction against a restricted area of the human epidermis (e.g., the skin on the tip of the finger) would result in damage to the epidermal membrane itself. Thus, there is a dual information-carrying nature that the interfacing bodily elements of biological systems have: they carry information about the conditions of the environment and the condition of the very system to which these interfacing elements belong. This dual information-carrying nature of interfacing bodily elements, I would like to argue, is

[10] Nanobots, however, are programmed to do specific tasks and, as such, they differ, at least currently, from, say, autonomous protein motors in biological systems (Linke et al., 2020).

[11] The theoretical beginnings of the current research programme of self-replicating machines may be traced back to Von Neumann's (1966) logical models of self-reproducing automata.

[12] Xenobots (Kriegman et al., 2020), algorithmically designed collections of frog skin and heart cells that have the capacity to heal themselves, may be a possible exception to this.

a second necessary feature for self-concern in biological systems; self-concern requires a structure through which multiple sources of environmental feedback are integrated and causally related to the condition of the system.

This feature places some important constraints on the materiality of embodied agents, a constraint that can be read-off of direct observation on how such interfacing bodily elements are implemented in biological systems. One observation is that the kinds of membranes that separate a biological system from its environment are deformable; their structure is responsive to changes in applied external and internal forces (tension, friction, and stress). Deformation of a material substrate not only implies a change in environmental conditions (forces applied) but also change in the structural condition of the deformed membrane and the system that possesses such membrane.

Another observation is that the kinds of membranes and interface components found in livings systems often exhibit material elasticity (i.e., having low elastic modulus). This condition is tightly (yet contingently) related to the property of deformability; after structural deformation due to change in environmental conditions (e.g., brief exertion of compressive force), materials that things like cell membranes or epidermal membranes are made of allowing for a system to return to its original structure. In other words, the materials allow for recovery from deformation. This is relevant for any system that uses information about the environment to guide its behaviour because if such a system not only detects deformation but also how long it takes to elastically recover from deformation, that system can use the time difference between deformation and recovery to direct its behaviour and measure the states of its own responsiveness, something that may be informative about damage; whereas a short (or regular) recovery time suggests that the condition of the material itself is suited for optimal performance, a long (or irregular) recovery time suggests that the material has perhaps a defect or structurally damaged and unsuitable for optimal performance.

A third and last observation regarding the kinds of membranes and interface components that separate a biological system from its environment is that they exhibit local transmission of causal effects. This is just to say that affecting one concentrated area in a membrane or other interface component is likely to also affect adjacent areas. This property is, of course, related to that of being deformable and elastic. For instance, dropping a weight on a taught sheet of rubber stretched across a frame does not only deform the area of the sheet where the weight meets the rubber but also increases the tension on the areas surrounding weight. Similarly, exerting mechanical force on a biological membrane does not only deform the point of contact but also affects the adjacent areas. This property of local transmission is important for biological systems, because it allows for information about the environment and the self to be distributed locally, and, hence, contributes to the biological system's ability to use different sources of information to monitor itself (its own current bodily conditions). To be sure, the property of local transmission that accompanies both elasticity and deformability suggests that any neat separation of exteroceptive and interoceptive (and proprioceptive) information may be an artificial one (see Gibson, 1966 for a similar remark about exteroception and proprioception).[13]

These three material properties characteristic of dual information-carrying interfacing bodily elements imply something crucial about the structure of an embodied AI if its structure is to support the emergence of self-concern, namely, it is (partly) due to the fact that biological systems are composed of biotic structures made of materials that are largely deformable, elastic, and transmit local causal effects that such systems are subject to damage that occurs on the timescale of living systems. This timescale of damage can be contrasted to the timescale of damage that would be sustained by a structure composed of non-deformable, rigid materials. Many of our longest-standing architectural constructions (e.g., pyramids and the Colosseum of Rome) have been constructed of rigid materials such as limestone and cement, and our more recent architectural constructions add the strength of steel. The durability of these materials is related to their rigidity (e.g., their high tensile strength and compressive strength). The kind of damage that they sustain takes the form of long-timescale processes of metal corrosion, aggregate expansion, and calcium and lime leaching, to name a few. Because of the deformable, elastic, and local transmission properties of materials that largely make up living bodies, the damage that things like cell membranes sustain might take the form of post elastic limit micro-tears or enzymatic decomposition (e.g., being digested by an amoeba). Since such damage is detectable and occurs at fast timescales (i.e., it is not due to being exposed to constant environmental conditions) it allows living systems to behave in ways that minimise future damage, fleeing the situation if motile, or nutating towards better conditions if sessile.[14]

Material properties that underwrite the ability of interfacing bodily elements to carry information about the world and the system itself are largely captured with new technologies of the rapidly developing field of soft robotics (Hawkes et al., 2017; Booth et al., 2018; Shih et al., 2019, 2020; Thuruthel et al., 2021; Hardman et al., 2022). Whereas traditional robotics, focusing upon task precision and strength in controlled environments,

[13]Although I have focused on membranes and their properties that are typical to living systems that exhibit dual information carrying, having a membrane is not a necessary requirement for dual information carrying to be physically instantiated. As long as there are distributed sensing/actuation devices that can harness and leverage spatially relative information, dual information carrying can be instantiated by a network of elements that are not, as a whole, enveloped by a single membrane. In such cases, the boundary that is instantiated is one of dynamically coupled causal influence. For example, a swarm of robots need not itself have a membrane to exhibit dual information carrying. All that is required is that the behaviour of each component element of the swarm is a function of the sensing/actuation parameters of the others. In such a case, the coupled swarm will exhibit dual information carrying across its network of distributed swarm members, each with their own sensing/actuation devices that contribute to an emerging dynamically bounded system.

[14]This is not to suggest that living systems are not subject to slow and gradual decomposing *via* oxidation reactions; they are if they are aerobic respiring systems or they live in oxygen-rich environments. Such damage, however, being the result of a constant environmental condition, fails to be something that an organism can escape from; it is merely a condition on aerobic life. As such, although complex organisms like us are aware of the slow damage that oxidation causes, it is not something that we are concerned with given it is a condition on our aerobic life.

have used (and still use!) rigid metal links (i.e., joints) and electric non-distributed actuators, soft robotics focuses on the adaptability of robots to the real world and complex environments, and, as such, uses synthetic compliant materials that can undergo deformation along with distributed actuators. In addition to this, conductive piezoresistive strain sensor fibres have recently been incorporated into the self-healing deformable material, allowing for the control system to sense damage (Georgopoulou et al., 2021) and proprioceptive feedback (Truby et al., 2020).[15] This feature seems to accurately map onto the material properties (deformability, elasticity, and local transmission of causal effects), which, as I have argued, are required for implementing interfacing bodily elements that carry information about both the environment and the agent. This being said, were such soft robots able to measure their own energetic conditions and actively (or proactively) engage in energy-sourcing and material exchange—the kind which we have already seen is suggested by controlled system-environment energy traffic, such robots would indeed be one step closer to the realisation of agents with self-concern (cf. Man and Damasio, 2019).

Let us now move on to the notion of a hierarchically structured organisation; the third and last feature that I will argue is necessary for concern in biological systems.

## Common Fate of Hierarchically Structured Systems

Living systems are self-organising complex adaptive systems. This means that they can maintain themselves in states that are far from thermodynamic equilibrium, temporarily avoiding the dissipation that inevitably results from the tendency for entropy to increase (Nicolis and Prigogine, 1977; Friston, 2012, 2019; Demirel, 2014). One central property of complex systems is that they are hierarchically organised (Simon, 1962). We may understand the notion of a hierarchical system in terms of "containment" where a larger system contains smaller subsystems nested within it (McShea, 2012). This is to say that such systems contain "interrelated subsystems, each of the latter being, in turn, hierarchic in structure until we reach some lowest level of elementary system" (Simon, 1962, p. 468).[16] Multicellular organisms, for example, contain living organs and tissues that, in turn, contain living cells. Although the hierarchical organisation of complex systems more generally may be construed in terms of intensity of interactions (i.e., who interacts with whom and how often), the form that the intensity of interaction takes in the hierarchical organisation of both biological and physical systems is that of "relative spatial propinquity" (Simon, 1962, p. 469).

Hierarchical organisation implies differences in both relative timescales and behavioural constraints; faster timescale behaviour of nested subsystems is constrained by slower timescale behaviour of systems in which they are nested. Roughly, system Y is constrained by another system, X, just in case the latter's features act as order parameters for the former, reducing the degrees of freedom of Y.[17] As such, we may say that the faster timescale dynamics of some nested system is "enslaved" to the slower global dynamics of the nesting system, making the former subordinate to the latter (Haken, 1985). For instance, the homeostatic condition of a liver (roughly, a collection of differentiated cells and biochemical cell interactions) may largely constrain the homeostatic state of any individual cell contained in the collection, and, simultaneously, the individual cells contribute to the maintenance of the liver that constrains them. Crucially, the kind of constraint relations that are governed by a biological system's homeostatic imperatives, unlike those in systems that do not answer to system-wide self-maintenance, means that "certain differences in the part have an informational effect upon the larger unit, and vice versa" (Bateson, 1972, p. 324). This is just to say that for both nested and nesting biological systems, a cause for homeostatic compensation at either of their respective levels is semantic information (Kolchinsky and Wolpert, 2018).

How is hierarchical organisation related to self-concern? To answer this question, let us first consider the fact that in multicellular biological systems every nested subordinate (bounded) system within a larger superordinate system is, itself, adaptive and "participates in its own self-maintenance, sensing and signalling the state of its life process" (Man and Damasio, 2019, p. 447). This suggests that self-concern is not merely a property of a superordinate system, but that it is something that arises at each nested level of biological organisation to varying degrees. Just as organisms adaptively respond to external environmental challenges (e.g., stresses) they face to avoid dyshomeostasis, organs and tissues modify their dynamics in ways that are adaptive to the challenging conditions of their body environment; similarly, bodily cells *via* variable gene expression biochemically respond to the challenges of their organ or tissue environments (Pezzulo and Levin, 2016; Levin, 2019).[18]

What about prokaryotic organisms that do not have organelles and yet still exhibit self-concern? Although elements within prokaryotes (e.g., microtubules and actin filaments of the cytoskeleton, etc.) may not exhibit self-concern themselves, something that follows from the fact that such parts fail to satisfy two necessary conditions for the self-concern proposed above, it is the dynamics occurring across different scales constrained by the temporal-spatial organisation of such elements that contribute to self-concern at the level of the organism. Importantly, these element dynamics may be characterised as

[15] The "Octobot" (Wehner et al., 2016) is a striking example of an entirely soft robot that contains no electronics. Much like signal processing in slime mould, the signal processing in the Octobot occurs *via* oscillations and fluid transport (microfluidic logic). Locomotion, signal processing, and decomposition of onboard fuel supply are closely connected!

[16] Of course, the identification of a lowest level is something that is, itself, an open question of scientific investigation and a matter of contention. However, for the purposes of this article, we can set the question of what such a level might be aside.

[17] In synergetics (Haken, 1985), an "order parameter" refers to a measure of a global system's slow macroscale dynamics that determines the fast dynamics of microscale component systems by reduction of the degrees of freedom of the latter.

[18] Parisi and Petrosino (2010) were early to stress how robotics could be instrumental for understanding how organisms adapt to both "external worlds" and "internal worlds." Conversely but in a similar spirit, I am suggesting that understanding how hierarchically nested/nesting biological systems adapt simultaneously to both internal and external worlds is crucial for the development of concern in biologically inspired robots.

serving a common fate; it is by contributing to the continued homeostatic maintenance of the whole organism that each element brings about the continued production of itself *via* the material turnaround that is insured by the continued functioning of the organism.

With these considerations in mind, I would like to suggest that self-concern is a system-level property that rests on there being a common fate of both the hierarchically nested parts and the nesting (i.e., constraining) system.[19] Without some degree of common fate being respected at relatively lower levels of hierarchical organisation, self-concern at higher levels is not possible. Imagine the rather brutal example that tomorrow every cell in your body would begin competing for resources with one another; this would be tantamount to each cell behaving as a unicellular organism (ignoring the presence of biofilms for the sake of argument!). Although each of these cells might exhibit self-concern in much the same manner that prokaryotes do, self-concern would fail to arise at the level of either the organ or the organism. Each cell would fail to contribute to the self-maintenance of your organs (or you at the organismic level) because of the absence of common cell fate; the viability of one cell in such a brutal case may indeed be largely dependent on bringing about conditions that would adversely affect the homeostasis of all others. Thus, self-concern in biological systems at the level of nesting (constraining) system requires that nested systems share some degree of common fate with nesting systems.

If the common fate of hierarchically organised systems is a requirement on biological self-concern, and such concern is a requirement on biological intelligent behaviour, then using the biological case as a model for agents, it seems that some degree of hierarchically organised systems exhibiting common fate will be a necessary feature for embodied AI if something even slightly akin to basic biological intelligence can be exhibited by AI. How can the notion of the common fate of hierarchically organised systems inform embodied AI more specifically?[20] One very general suggestion is to explore the endosymbiotic relationship and design artificial endosymbiotic systems. The idea is to somehow elicit self-concern at the level of global nesting-system from both (a) the interactivity of the global system and a simple nested system, which have both been programmed to emulate a form of mutualism (i.e., to use and contribute to one another's continued self-maintenance) and (b) the agent-environment energy (metabolic) traffic that the global system is forced to engage in to keep the endosymbiont and itself functioning. Treating the endosymbiont agent as the lower, faster scale subordinate system and the host agent as the slower scale superordinate system, the hierarchical organisation of these systems sharing a degree of common fate, amongst some of the other features that we have already touched upon, may support a self-organising autonomous system, a system that by bootstrapping[21] can cast away its scaffolding, replacing the self-maintenance tasks extrinsically put in place by designers with tasks that arise (i.e., are discovered) from interactions between the parts, tasks that reflect an intrinsic concern on the part of the whole system for its own self-maintenance.

Recent developments in swarm robotics may prove to be a promising avenue for testing this hypothesis (see Christensen et al., 2008; Liu and Winfield, 2009 for symbiotic-inspired robot swarms). Here, interaction among robots has taken the form of continuous (Bezzo et al., 2014) or pulse-coupled oscillatory signals (Barcis et al., 2019), whose alignment of oscillatory frequency and/or phase alignment allow members of a swarm to behave synchronously (Schranz et al., 2020). If swarm robots were to implement the kind of hierarchical common fate dynamics, which, as I have argued, is required for self-concern, then the interaction among them must not only take the form of information-sharing but must also involve behavioural interaction by exerting reciprocal mechanical forces that have consequences on the continued functioning of both the nested and nesting parts.

Another research area that may lend itself to the development of agents with common fate at various hierarchical levels of an organisation is that of hybrid associations composed of both machines and microorganisms (Ieropoulos et al., 2005; Philamore et al., 2016; Tsompanas et al., 2021). For example, anaerobic *anodophile* bacteria, when used in a microbial fuel cell (MFC), will transfer electrons to the MFC's electron-accepting anode electrode, which in turn supplies an MFC housing robot with electrical energy (Habermann and Pommer, 1991; Ieropolous et al., 2004).[22] If such an MFC housing robot behaves so as to remain in (or return to) environments that are rich in kinds of substances that MFC-inhabiting bacteria can metabolise (e.g., sulphate, acetate, glucose), then its doing so mutually supports the survival of the bacteria and its own continued energetic functioning. Further development of this kind of (biotic/abiotic) artificial symbiosis between nested bacteria and a nesting machine, if what has been argued here is correct, may be one crucial method of bringing about self-concern in AI, namely, bacteria's self-concern may be the source of the emergence of self-concern in the larger hierarchically organised bacteria-machine system.

One important takeaway from this section is this: the fact that self-concern has not yet been implemented machines (Man and Damasio, 2019, p. 447) does not suggest that such self-concern cannot find expression in machines or organism-machine associations. If self-concern is a requirement for basic intelligence across the board, then we should expect a concentrated effort on the part of future research to develop hierarchically organised systems that implement self-concern across scales.

[19]The notion of common fate was originally introduced by Wilson and Sober (1989) in the context of providing a characterisation of biological individuals.

[20]It should be emphasised that the kind of nested hierarchical organisation of elements with common fate that I am arguing about is fundamental to (self) concerned systems involved in concrete physical implementation rather than merely a hierarchically organised network (architecture) that plays the role of a control system. Many thanks to Christian Oettmeier for pushing me to clarify this point.

[21]The term "bootstrapping" is used in a general manner here to denote a "process that automatically increases in complexity" (Moses and Chirikjian, 2020, p. 9) rather than the specific notion of bootstrapping deployed in statistics.

[22]MFCs are transducers that are given a biochemical energy source and convert it to electrical energy; they can, thus, power robots in which they are located.

## WHEN ANTICIPATORY DYNAMICS ANSWER TO SELF-CONCERN

With three necessary features of biological self-concern on hand, I now want to consider the relationship between self-concern and intelligence by way of a particular capacity that is central to cognition: anticipatory behaviour. My argument in this final section will take the following form: anticipatory behaviour is often recognised as indicative of intelligence; an important aspect of intelligent anticipatory behaviour (in contrast to mere anticipatory behaviour) is that it is possible for such behaviour to answer to a system's long-term homeostasis; a system's behaviour that answers to its long-term homeostatic stability is behaviour that answers to that system's self-concern; therefore, an important aspect of intelligent anticipatory behaviour is that it is possible for such behaviour to answer to self-concern. If one effective way of bringing embodied AI closer to emulating the intelligence found in even the most basal of living systems is to use aspects of biological intelligence to inform AI design, then designing agents, the anticipatory behaviour of which could answer to self-concern, may be one manner of bringing AI closer to emulating the kind of intelligence observed in the biological world.

It has been widely acknowledged that one way of expressing intelligence in biological systems is by an exhibition of anticipatory behaviour (Bartlett, 1932; Craik, 1943; Piaget, 1970; Neisser, 1976; Drescher, 1991; Arbib, 1992; Riegler, 2001; Grush, 2004; Castelfranchi, 2005; Lyon, 2006; Bar, 2007; Pezzulo, 2008; Bickhard, 2016; Nasuto and Hayashi, 2016; Levin, 2019; Kiverstein and Sims, 2021; Sims, 2021). Anticipatory behaviour may be generally characterised as a behaviour that allows for a system to respond to yet-to-be encountered changes in external or internal environmental states as a function of prior states that the system has encountered. This characterisation throws light on one reason why anticipatory behaviour is considered as an expression of intelligence: it involves deploying some form of memory and learning (or model acquisition more generally) to bias behaviour towards a system-preferred outcome; hence, it involves some form of information processing. Such behaviour is orchestrated in a manner that is dependent on internal states that have a certain degree of independence or detachment from current streams of sensory information (Pezzulo, 2008). Internal states may be generally construed as constituting a system's "internal model" that captures environmental dynamics and the effects of its actions on its environment (Neisser, 1976; Rosen, 1985/2012; Pezzulo, 2008; Friston, 2012; Pezzulo and Levin, 2016; Schulkin and Sterling, 2019).

Anticipatory behaviour, when all goes well, delivers preferred behavioural outcomes. Such outcomes are relative to physiological states that a system should visit given both its phenotype and the form of metabolic redox machinery that its phenotype serves (i.e., kinds of donors and acceptors a system implements to fuel proton chain reactions to drive catabolism of ATP). A preferred behavioural outcome for *E. coli*, which metabolises glucose, is encountering high concentrations of glucose in its environment. On the other hand, a preferred behavioural outcome for a sun-loving plant such as *Portulaca oleracea*, which requires photosynthetic light to effectively convert $H_2O$ and $CO_2$ into sugars, is encountering the presence of photosynthetic light. It is because preferences exhibit a high degree of stability and can act as reference points for long-term homeostatic supporting behaviour in the face of environmental flux (i.e., set-point values) that preferences themselves can be understood as (partly) constitutive of a system's internal model. Recall that long-term homeostasis is an intrinsic existential goal in all biological systems; as such, a system's behaviour that answers to its long-term homeostasis is a behaviour that answers to that system's self-concern (Jonas, 1966/2001). It is, thus, the ability for anticipatory behaviour to be driven by a system's concern for its own continued stability that qualifies such behaviour as intelligent at its most fundamental level. For example, a hypothetical system that only behaves anticipatorily, bringing about outcomes that are irrelevant to its continued survival, would certainly fail to survive very long; this would be because of the fact that no behaviour comes without some metabolic cost, and that regularly engaging in anticipatory behaviour that brings about expected sensory or behavioural outcomes is consistent with regularly bringing about maladaptive conditions. The regular and repeated proactive jumping of a mouse to the exact next location where a snake predator will strike falls short on any account of being an example of intelligent behaviour. Such a mouse, despite the accuracy of its predictions, is a dead mouse.

There is an objection waiting in the wings, which may be posed as follows: certainly, an elaborately planned suicide can be an exhibition of intelligence, and such a plan neither answers to long-term homeostasis nor self-concern on the part of the organism! This counterexample can only go through, however, if the claim was being made that all intelligent anticipatory behaviour must answer to a system's self-concern. The claim that I am making, however, is only that in order for some anticipatory behaviour in a biological system to qualify as intelligent, it must be possible for such behaviour to answer to self-concern grounded in the continued long-term homeostasis of that system.[23] In other words, if a system's anticipatory behaviour could not, in principle, be influenced by its homeostatic norms, then such behaviour, although it is anticipatory, fails to exhibit the form of intelligence that is typical of biological anticipatory behaviour. Hence, even though an elaborately planned suicide results in loss of life and dyshomeostasis, it is the type of behaviour that could, in fact, answer to the maintenance of long-term homeostatic stability. On such an occasion, however, it simply fails to do so.

There have been a number of recent cognitive theories and computational models that have taken into account the role of anticipation for intelligent behaviour and that have been used to inform cognitive robotic technologies (see Nasuto and Hayashi, 2016 for an overview).[24]One increasingly popular process

[23]I am concerned here with nomological possibility and not logical possibility.

[24]The implementation of anticipatory dynamics, of course, is not new. In optimal control theory, forward models (i.e., internal models that generate predictions of the sensory consequences of motor commands) have been used at least since Jordan and Rumelhart (1992).

theory that has been used to formulate specific anticipation-driven cognitive architectures is active inference (Friston et al., 2009; Pezzulo et al., 2015; Morville et al., 2018; Baltieri and Buckley, 2019; Corcoran et al., 2020; Millidge, 2020). This theory, which was originally applied to brain dynamics, casts perception, action, and learning in terms of a Bayesian inference problem-and-error correction. In active inference, agents are endowed with prior beliefs that they will frequent some sensory states more than others, and these priors reflect an agent's homeostatic range. Unlike rewards in classical reinforcement learning, which are received from the environment, priors in active inference agents are internal states of the agent that can remain stable across environments and, hence, in this sense may be construed as intrinsic to the agent (Friston et al., 2015). Similarly, Keramati and Gutkin's (2014) homeostatic reinforcement learning model aims to make sense of how rewarding behavioural outcome values are computed as a function of internal states and estimate dyshomeostasis reduction in an outcome.

Intelligent anticipatory behaviour in AI requires that it is possible for behaviour to answer to an agent's own intrinsic homeostatic goals; such behaviour is grounded in self-concern. Whether or not anticipatory architectures such as active inference (or homeostatic reinforcement learning for that matter) will bring embodied AI closer to intrinsically emerging normativity is dependent on the ability of such architectures to provide agents with means to go beyond the values that designers initially endow them with. In the case of active inference, this will likely include providing agents with a means for discovering conjugate priors (i.e., hyperpriors) (see Sajid et al., 2021) within real world environments that pose concrete threats to the continued self-maintenance of the agent. This kind of adaptive plastic reshaping of priors (i.e., value) may be roughly conceptualised as a form of accelerated evolution where a single agent is seen as an evolving lineage subject to open-ended learning (Standish, 2003; Stanley et al., 2017). By equipping an agent with the necessary means to learn the normative parameters that define its continued self-maintenance in real-world environments, the anticipatory behaviour of such an embodied agent is poised to answer to its own self-concern. What more is required? If what I have presented above is accurate, then focusing on the further development of (1) controlled system-environment energy traffic, (2) dual information-carrying nature of interfacing bodily elements, and (3) highly integrated, hierarchically structured systems sharing a common fate will be necessary to bring the current AI, the various architectures of which are already capable of generating accurate estimations of yet-to-be encountered states, much closer to emulating the kind of intelligent behaviour exhibited by biological systems, a behaviour that is grounded in self-concern.

## CONCLUSION

In this article, I have argued for the centrality of concern in biological systems even for the most basal expressions of intelligent behaviour. I argued that if this is the case, then there is good reason to think that if the intelligence of an agent is to ever be comparable to the intelligence of a biological system (human or otherwise), then it will require the presence of some form of self-concern. I have described three necessary features underpinning the concern in biological systems that can be used to inform the development of functionally similar features in embodied AI. Lastly, I have argued that although anticipation may be seen as central to intelligent behaviour, it is only anticipatory behaviour that could answer to the intrinsic norms of the system and, thus, be subject to self-concern which is a clear exhibition of intelligence.

To close, let us look at a few questions that highlight some of the complex issues that arise when considering the possible implementation of self-concern in embodied AI. Although addressing these questions falls beyond the scope of this article, I would like to stress the importance of considering them.

Let us assume that we have managed to implement self-concern in an embodied AI. This agent will be very unlike any of those that we currently interact with in at least one central manner; a self-concerned embodied AI agent will do what it can to remain in states that are consistent with its continued functioning/existence. This hypothetical case provides us with the opportunity, in closing, to raise a few important questions on potential challenges and ethical issues regarding self-concerned AI.

One challenge is to design self-concerned embodied AI agents in a way that avoids deceiving us (a form of explanatory opacity *via* transmission of misinformation). This will be crucial for such AI agents given the fact that if they exhibit the kind of basic biological intelligence that I have argued arises with self-concern, such agents may, in fact, disguise their immediate goals (deception) in order to satisfy their long-term goal of continued self-maintenance, not unlike many living organisms do (e.g., mimicry, feigning death, etc.). A further challenge is this: one can easily imagine a case in which such a self-concerned AI agent's continued functioning might be incompatible (or evolves over time to be incompatible) with the continued well-being of a human (or multiple humans) (e.g., competition for common resources, etc.). How can this scenario be avoided *via* precautionary design efforts without jeopardising the very aim of self-concern in embodied AI? A related ethical question arises when considering that an embodied AI agent that implements self-concern need not be one that is self-aware or conscious. Should the fact that we are aware that such an AI agent has an intrinsic concern about its own continued functioning be enough for us to be obligated to avoid impeding its intrinsic goals? If so, what are the situations in which we can renege on such obligations?

One important question that we are left with is whether self-concern in embodied AI is something that could even be recognisable by us? Although it might take the form of machines that flexibly and adaptively behave in ways that allow them to avoid their own "machine death," the kind of behaviour that will be driven by machine self-concern is likely to be very different

from that which we can readily identify in living systems. Perhaps most interesting here is the prospect of learning more about the aspects of self-concern that are particular to humans from our attempts at making sense of self-concern in machines. Whatever the case, although the emergence of concern in AI may very well depend on us, once up and running, the intelligence that such systems exhibit will be directed at completing tasks that are (largely) their own.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## FUNDING

This research has been supported by the Alexander Von Humboldt Foundation.

## ACKNOWLEDGMENTS

I would like to thank the following people for the invaluable input and support: Tobias Schlicht and his research group, Michael Levin, Pamela Lyon, Giovanni Pezzulo, Christina Oettmeier, Nick Brancazio, Carrie Figdor, Kate Nave, Ian Mason, and Ben Little. I would also like to express my gratitude to the two reviewers for their generous feedback and helpful comments.

## REFERENCES

Arbib, M. A. (1992). "Schema theory," in *Encyclopedia of Artificial Intelligence*, *2nd Edn*, eds S. Shapiro (Chichester: Wiley), 1427–1443.

Ashby, W. R. (1952). *Design for a Brain, 1st Edn*. London: Chapman and Hall.

Baltieri, M., and Buckley, C. L. (2019). "Nonmodular Architectures of Cognitive Systems based on Active Inference," in *2019International Joint Conference on Neural Networks (IJCNN),* 1–8.

Baluška, F., and Levin, M. (2016). On having no head: cognition throughout biological systems. *Front. Psychol.* 7, 902. doi: 10.3389/fpsyg.2016.00902

Baluška, F., and Mancuso, S. (2009). Deep evolutionary origins of neurobiology: Turning the essence of "neural" upside-down. *Commun. Integrat. Biol.* 2, 60–65. doi: 10.4161/cib.2.1.7620

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends Cogn. Sci.* 11, 280–289. doi: 10.1016/j.tics.2007.05.005

Barcis, A., Barcis, M., and Bettstetter, C. (2019). "Robots that sync and swarm: a proof of concept in ROS 2," in *Proceedings of the International Symposium on Multi-Robot and Multi-Agent Systems* (New Brunswick, NJ: IEEE), 98–104.

Barrett, N. (2019). On the nature and origins of cognition as a form of motivated activity. *Adapt Behav.* 28, 89–103. doi: 10.1177/1059712318824325

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press: Cambridge.

Bateson, G. (1972). *Steps to an Ecology of Mind*. New York, NY: Chandler Publishing Co., Balantine Books, Random House.

Bechtel, W., and Bich, L. (2021). Grounding cognition: heterarchical control mechanisms in biology. *Philos. Trans. R. Soc. B* 376, 20190751. doi: 10.1098/rstb.2019.0751

Ben Jacob, E., Shapira, Y., and Tauber, A. (2006). Seeking the Foundation of Cognition in Bacteria: From Schödinger's negative entropy to latent information. *Phys. Stat. Mech. Appl.* 369, 495–524. doi: 10.1016/j.physa.2005.05.096

Berger, M. (2016). *Nanotechnology: The Future Is Tiny*. Cambridge: Royal Society of Chemistry.

Berridge, K. C., and Robinson, T. E. (2003). Parsing reward. *Trends Neurosci.* 26, 507–513. doi: 10.1016/S0166-2236(03)00233-9

Bezzo, N., Cruz, P. J., Sorrentino, F., and Fierro, R. (2014). Decentralized identification and control of networks of coupled mobile platforms through adaptive synchronization of chaos. *Phys D Nonlin Phenom.* 267, 94–103. doi: 10.1016/j.physd.2013.08.012

Bickhard, M. H. (2016). "The anticipatory brain: two approaches," in *Fundamental Issues of Artificial Intelligence, Vol 326*, ed V. C. Müller (Berlin: Springer Synthese Library), 259–281.

Bongard, J., and Levin, M. (2021). Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors inLight of the Modern Science of Machine Behavior. *Front. Ecol. Evol.* 9, 650726. doi: 10.3389/fevo.2021.650726

Booth, J. W., Shah, D., Case, J. C., White, E. L., Yuen, M. C., Cyr-Choiniere, O., Kramer-Bottiglio, R. (2018).Omniskins: Robotic skins that turn inanimate objects into multi- functional robots. *Sci. Robot.* 3, eaat1853. doi: 10.1126/scirobotics.aat1853

Boussard, A., Fessel, A. Oettmeier, C., Briard, L., Döereiner, H. G., and Dussutour, A. (2021). Adaptive behaviour and learning in slime moulds: the role of oscillations*Phil. Trans. R. Soc. B* 376, 20190757. doi: 10.1098/rstb.2019.0757

Calvo, P., Gagliano, M., Souza, G. M., and Trewavas, A. (2020). Plants are intelligent, here's how. *Ann. Bot.* 125, 11–28. doi: 10.1093/aob/mcz155

Cassirer, E. (1981). *Kant's Life and Thought*. New Haven: Yale University Press.

Castelfranchi, C. (2005). "Mind as an anticipatory device: for a theory of expectations," in *BVAI 2005. LNCS*, *Vol. 3704*, eds M. De Gregorio, V. Di Maio, M. Frucci, and C. Musio (Heidelberg: Springer), 258–276.

Christensen, A., O'Grady, R., and Dorigo, M. (2008). Swarmorphscript: a language for arbitrary morphology generation in self-assembling robots. *Swarm Intell.* 2, 143–165. doi: 10.1007/s11721-008-0012-6

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together*. Cambridge, MA: MIT Press.

Conant, R., and Ashby, W. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97.

Cooper, S. T., and McNeil, P. L. (2015). Membrane repair: mechanisms and pathophysiology. *Physiol. Rev.* 95, 1205–1240. doi: 10.1152/physrev.00037.2014

Corcoran, A. W., Pezzulo, G., and Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biol. Philos.* 35, 32. doi: 10.1007/s10539-020-09746-2

Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex. 1st Edn, 2 vols*. London: John Murray.

Demirel, Y. (2014). *Nonequilibrium Thermodynamics: Transport and Rate Processes in Physical, Chemical and Biological Systems. Third Edition*. Amsterdam: Elsevier.

Drescher, G. L. (1991). *Made-Up Minds: A Constructivist Approach to Artificial Intelligence*. Cambridge, MA: MIT Press.

Dufrêne, Y. F., and Persat, A. (2020). Mechanomicrobiology: how bacteria sense and respond to forces. *Nat. Rev. Microbiol.* 18, 227–240. doi: 10.1038/s41579-019-0314-2

Dupeyroux, J., Diperi, J., Boyron, M., Viollet, S., and Serres, J. (2017). "A bio-inspired celestial compass applied to an ant-inspired robot for autonomous navigation," in *2017 European Conference on Mobile Robots (ECMR)*. IEEE, 1–6.

Fraikina, G. Y., Strakhovskayaa, M. G., Belenikinaa, N. S., and Rubina, A. B. (2015). Bacterial photosensory proteins: Regulatory functions and optogenetic applications. *Mikrobiologiya* 84, 391–403. doi: 10.1134/S0026261715040086

Friston, K. (2019). A free energy principle for a particular physics. *Neurons Cogn.* arXiv1906.10184.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE* 4:e6421. doi: 10.1371/journal.pone.0006421

Gagliano, M., Vyazovskiy, V. V., Borbély, A. A., Grimonprez, M., and Depczynski, M. (2016). Learning by association in plants. *Sci. Rep.* 6, 38427. doi: 10.1038/srep38427

Galetti, G., Rossi, S., Caffarra, C., Gerboles, A., and Miragoli, M. (2019). "Innovation in nanomedicine and engineered nanomaterials for therapeutic purposes," in *Micro and Nano Technologies, Exposure to Engineered Nanomaterials in the Environment,* eds N. Marmiroli, J. C. White, J. Song (Amsterdam: Elsevier).

Gánti, T. (2003). *The Principles of Life*. Oxford: Oxford University Press.

Georgopoulou, A., Bosman, A. W., Brancart, J., Vanderborght, B., and Clemens, F. (2021). Supramolecular self-healing sensor fiber composites for damage detection in piezoresistive electronic skin for soft robots. *Polymers* 13, 2983. doi: 10.3390/polym13172983

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton-Mifflin.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge, UK: Cambridge University Press.

Godfrey-Smith, P. (2016). Individuality, subjectivity, and minimal cognition. *Biol. Philos.* 31, 775–796. doi: 10.1007/s10539-016-9543-1

Goodwin, B. C. (1978). A cognitive view of biological processes. *F. Social Biol. Struct.* 1, 111–125. doi: 10.1016/S0140-1750(78)80001-3

Goodwin, B. C. (1994). *How the Leopard Changed its Spots: The Evolution of Complexity*. Princeton: NJ. Princeton University Press.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–442. doi: 10.1017/S0140525X04000093

Habermann, W., and Pommer, E-H. (1991). Biological fuel cells with sulphide storage capacity. *J. Appl. Microbiol. Biotechnol.* 35, 128-133. doi: 10.1007/BF00180650

Haken, H. (1985). Synergetics: an interdisciplinary approach to self-organization. *Geoform* 16, 205–211. doi: 10.1016/0016-7185(85)90029-6

Hanson, A. (2021). Spontaneous electrical low-frequency oscillations: a possible role in *Hydra* and all living systems. *Philos. Trans. R. Soc. B* 376: 2019076320190763. doi: 10.1098/rstb.2019.0763

Hardman, D., Thuruthel, T. G., and Iida, F. (2022). Self-healing ionic gelatin/glycerol hydrogels for strain sensing applications. *NPG Asia Mater.* 14, 1–13. doi: 10.1038/s41427-022-00357-9

Hawkes, E. W., Blumenschein, L. H., Greer, J. D., Okamura, A. M. (2017). A soft robot that navigates its environment through growth. *Sci. Robot.* 2, eaan3028. doi: 10.1126/scirobotics.aan3028

Hellingwerf, K. J. (2005). Bacterial observations: a rudimentary form of intelligence? *Trends Microbiol.* 13, 152–158. doi: 10.1016/j.tim.2005.02.001

Hull, C. L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. New York, NY: Appleton-Century-Croft.

Ieropolous, I., Melhuish, C., and Greenman, J. (2004). "Energetically autonomous robots," in *Proceedings of the Eighth Intelligent Autonomous Systems Conference* (IAS-*8), Amsterdam,* 128–135.

Ieropoulos, I., Greenman, J., Melhuish, C., and Hart, J. (2005). Comparative study of three types of microbial fuel cell. *Enzyme Microbial. Technol.* 37, 238–245. doi: 10.1016/j.enzmictec.2005.03.006

Jonas, H. (1966/2001). *The Phenomenon of Life:Toward a Philosophical Biology*. Evanston, IL: Northwestern University Press.

Jordan, M., and Rumelhart, D. E. (1992). Forward models: supervised learning with a distal teacher. *Cogn. Sci.* 16, 307–354. doi: 10.1207/s15516709cog1603_1

Kant, I. (1790/2007). *Critique of Judgement*. Transl. J. C. Meredith. Oxford: Oxford University Press.

Kauffman, S. (2000). *Investigations*. New York, NY: Oxford University Press.

Keramati, M., and Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward learning collection and physiological stability. *eLife* 3, 1–26. doi: 10.7554/eLife.04811.032

Kiverstein, J., and Sims, M. (2021). Is free-energy minimisation the mark of the cognitive? *Biol. Philos.* 36, 25. doi: 10.1007/s10539-021-09788-0

Kolchinsky, A., and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* 8, 20180041. doi: 10.1098/rsfs.2018.0041

Konidaris, G., and Barto, A. (2006). "An adaptive robot motivational system," in *From Animals to Animats 9: Proceedings of the 9th International Conference on Simulation of Adaptive Behavior* (Roma, Italy, SAB-06).

Kriegman, S., Blackiston, D., Levin, M., and Bongard. J. (2020). A scalable pipeline for designing reconfigurable organisms. *PNAS* 117, 1853–9. doi: 10.1073/pnas.1910837117

Lane, N. (2016). *The Vital Question: Why is Life the Way It Is*? London: Profile Books LTD.

Lara, B., Astorga, D., Mendoza-Bock, E., Pardo, M., Escobar, E., and Ciria, A. (2018). Embodied cognitive robotics and the learning of sensorimotor schemes. *Adap. Behav.* 26, 225–238. doi: 10.1177/1059712318780679

Lehman, J., and Stanley, K. O. (2011). "Novelty search and the problem with objectives," in *Genetic Programming Theory and Practice*, eds R. Riola, E. Vladislavleva, and J. H. Moore (Berlin: Springer), 37-56.

Levin, M. (2019). The computational boundary of a "self": developmental bioelectricity drives multicellularity and scale-free cognition. *Front. Psychol.* 10, 2688. doi: 10.3389/fpsyg.2019.02688

Linke, H., Höcker, B., Furuta, K., Forde, N., and Curmi, P. M. G. (2020). Synthetic biology approaches to dissecting linear motor protein function: towards the design and synthesis of artificial autonomous protein walkers. *Biophys. Rev.* 12, 1041–1054. doi: 10.1007/s12551-020-00717-1

Liu, W., and Winfield, A. (2009). "Implementation of an IR approach for autonomous docking in a self-configurable robotics system," in *Proceedings of Towards Autonomous Robotic Systems*.

Lyon, P. (2006). The biogenic approach to cognition. *Cogn. Process.* 7, 11–29. doi: 10.1007/s10339-005-0016-8

Lyon, P., Keijzer, F., Arendt, D., and Levin, M. (2021). Reframing cognition: getting down to biological basics. *Philos. Trans. R. Soc. B* 376, 20190750. doi: 10.1098/rstb.2019.0750

Macnab, R. M., and Koshland, D. E. (1972). The gradient-sensing mechanism in bacterialchemotaxis. *PNAS* 69, 2509–2512. doi: 10.1073/pnas.69.9.2509

Man, K., and Damasio, A. (2019). Homeostasis and soft robots in the design of feelings machines. *Nat. Mach. Intell.* 1, 446–452. doi: 10.1038/s42256-019-0103-7

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: the Realization of the Living.* Dordrecht: D. Reidel Publishing Company.

Maturana, H. R., and Varela, F. J. (1987). *The Tree of Knowledge.* Boston: New Science Library/Shambhala Publications.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine* 27, 12–14.

McShea, D. W. (2012). Upper-directed systems: a new approach to teleology in biology. *Biol. Philos.* 27, 663-684. doi: 10.1007/s10539-012-9326-2

Meincke, A. S. (2018). *Bio-Agency and the Possibility of Artificial Agents in: Philosophy of Science—Between the Natural Sciences, the Social Sciences, and the Humanities*. ed. by G. Schurz (Dordrecht: Springer, European Philosophy of Science Association Series).

Millidge, B. (2020). Deep active inference as variational policy gradients. *J. Math. Psychol.* 96, 102348. doi: 10.1016/j.jmp.2020.102348

Morville, T., Frison, T., Burdakov, D., Siebner, H., and Hulme, O. (2018). The homeostatic logic of reward. *bioRxiv Preprints* doi: 10.1101/242974

Moses, M., and Chirikjian, G. S. (2020). Robotic Self-Replication. *Annu. Rev. Control Robot. Autonom. Syst.* 3, 1–24. doi: 10.1146/annurev-control-071819-010010

Nakagaki, T., Yamada, H., and Tóth, A. (2000). Maze-solving by an amoeboid organism. *Nature* 407, 470. doi: 10.1038/35035159

Nasuto, S. J., and Hayashi, Y. (2016). Anticipation: beyond synthetic biology and cognitive robotics. *Biosystems*. 148, 22–31. doi: 10.1016/j.biosystems.2016.07.011

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. New York: Freeman.

Nicholson, D. J. (2020). "On being the right size, revisited: the problem with engineering metaphors in molecular biology," in *Philosophical Perspectives on the Engineering Approach in Biology: Living Machines?* eds S. Holm and M. Serbaneds (London: Routledge).

Nicolis, G., and Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems.* New York, NY: John Wiley.

Novoplansky, A. (2016). "Future perception in plants," in *Anticipation Across Disciplines,* ed M. Nadin (Berlin: Springer), 57–70.

Oudeyer, P., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1, 1–14. doi: 10.3389/neuro.12.006.2007

Parisi, D., and Petrosino, G. (2010). Robots that have emotions. *Adap. Behav.* 18, 453–469. doi: 10.1177/1059712310388528

Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation. *Minds Mach.* 18, 179–220. doi: 10.1007/s11023-008-9095-5

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., and Spivey, M. J. (2011). The mechanics of embodiment: a dialog on embodiment and computational modelling. *Front. Psychol.* 2, 5. doi: 10.3389/fpsyg.2011.00005

Pezzulo, G., and Levin, M (2016). Top-down models in biology: explanation and control of complex living systems above the molecular level. *J. R. Soc. Interface* 13, 20160555. doi: 10.1098/rsif.2016.0555

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol.* 134, 17–35. doi: 10.1016/j.pneurobio.2015.09.001

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. Cambridge, MA: The MIT Press.

Pfeifer, R., Lungarella, M., and Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science* 318, 1088–1093. doi: 10.1126/science.1145803

Philamore, H., Ieropoulos, I., Stinchcombe, A., and Rossiter, J. (2016). Toward energetically autonomous foraging soft robots. *Soft Robot.* 3, 186–197. doi: 10.1089/soro.2016.0020

Piaget, J. (1970). *Genetic Epistemology*. New York: Columbia University Press.

Pinto, D., and Mascher, T. (2016). (Actino) Bacterial "intelligence": using comparative genomics to unravel the information processing capacities of microbes. *Curr. Genet.* 62, 487–498. doi: 10.1007/s00294-016-0569-3

Reid, C. R., and Latty, T. (2016). Collective behaviour and swarm intelligence in slime moulds. *FEMS Microbiol. Rev.* 40, 798–806. doi: 10.1093/femsre/fuw033

Riegler, A. (2001). "The role of anticipation in cognition," in *Computing Anticipatory Systems*, eds D. M. Dubois (Melville: AIP Proceedings), 534–541.

Rosen, R. (1985/2012). *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Pergamon, Oxford.

Saigusa, T., Tero, A., Nakagaki, T. and Kuramoto, Y. (2008). Amoebae anticipate periodic events. *Phys. Rev. Lett.* 100, 18101. doi: 10.1103/PhysRevLett.100.018101

Sajid, N., Tigas, P., Zakharov, A., Fountas, Z., and Friston, K. (2021). Exploration and preference satisfaction trade-off in reward-free learning. *arXiv* preprint. *arXiv*:2106.04316.

Schranz, M., Umlauft, M., Sende, M., and Elmenreich, W. (2020). Swarm robotic behaviors and current applications. *Front. Robot. AI* 7, 36. doi: 10.3389/frobt.2020.00036

Schulkin, J., and Sterling, P. (2019). Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends Neurosci.* 42, 740–752. doi: 10.1016/j.tins.2019.07.010

Sejnowski, T. (2018). *The Deep Learning Revolution*. Cambridge, MA: MIT Press.

Service, R. F. (2016). Chemistry Nobel heralds age of molecular machines. *Science.* 354, 158–159. doi: 10.1126/science.354.6309.158

Shapiro, J. A. (2007). Bacteria are small but not stupid: cognition, natural genetic engineering and sociobacteriology. *Stud. Hist. Philos. Biol. Biomed. Sci.* 38, 807–819. doi: 10.1016/j.shpsc.2007.09.010

Shemesh, H., Arbiv, A., Gersani, M., Ovadia, O., and Novoplansky, A. (2010). The effects of nutrient dynamics on root patch choice. *PLoS ONE* 5, e10824. doi: 10.1371/journal.pone.0010824

Shih, B., Christianson, C., Gillespie, K., Lee, S., Mayeda, J., Huo, Z., Tolley, M. T. (2019). Design considerations for 3D printed, soft, multimaterial resistive sensors for soft robotics. *Front. Robot. AI* 6, 30. doi: 10.3389/frobt.2019.00030

Shih, B., Shah, D., Li, J., Thuruthel, T. G., Park, Y., Iida, F., et al. (2020). Electronic skins and machine learning for intelligent soft robots. *Sci. Robot.* 5, eaaz9239. doi: 10.1126/scirobotics.aaz9239

Simon, H. A. (1962). The architecture of complexity. *Proc. Am. Philos. Soc.* 106, 467–482.

Sims, M. (2021). A continuum of intentionality: linking the biogenic and anthropogenic approaches to cognition. *Biol. Philos*. 36:51. doi: 10.1007/s10539-021-09827-w

Standish, R. K. (2003). Open-ended artificial evolution. *Int. J. Comput. Intell. Appl.* 3, 167–175. doi: 10.1142/S1469026803000914

Stanley, K. O., Lehman, J., and Soros, L. (2017). *Open-Endedness: The Last Grand Challenge You've Never Heard of While Open-Endedness Could be a Force for Discovering Intelligence, it Could Also be a Component of AI itself.* Sebastopol, CA: O'Reilly Media, Inc.

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol. Behav.* 106, 5–15. doi: 10.1016/j.physbeh.2011.06.004

Sutton, R., and Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Thórisson, K. R., Bieger, J., Thorarensen, T., Sigurð*ardóttir, J. S., and Steunebrink, B. R. (2016). "Why artificial intelligence needs a task theory," in *Artificial General Intelligence. AGI 2016. Lecture Notes in Computer Science, vol 9782*, eds B. Steunebrink, P. Wang, B. Goertzel (Cham: Springer).

Thuruthel, T. G., Falotico, E., Beccai, L., and Iida, F. (2021). Machine learning techniques for soft robots. *Front. Robot. AI* 8, 205. doi: 10.3389/frobt.2021.726774

Trewavas, A. (2014). *Plant Behaviour and Intelligence*. Oxford, UK: Oxford University Press.

Truby, R. L., Santina, C., and Rus, D. (2020). Distributed Proprioception of 3D Configuration in Soft Sensorized Robots via Deep Learning. *IEEE Robot. Autom. Lett.* 5, 3299–3306. doi: 10.1109/LRA.2020.2976320

Tsompanas, M.-A., You J., Philamore, H., Rossiter, J., and Ieropoulos, I. (2021). Neural networks predicting microbial fuel cells output for soft robotics applications. *Front. Robot. AI* 8, 633414. doi: 10.3389/frobt.2021.633414

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Van Duijn, M., Keijzer, F. A., and Franken, D. (2006). Principles of minimal cognition: casting cognition as sensorimotor coordination. *Adapt. Behav.* 14, 157–170. doi: 10.1177/105971230601400207

Von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. Champaign: Univ. Ill. Press.

Webb, B., and Scutt, T. (2000). A simple latency-dependent spiking-neuron model of cricket phonotaxis. *Biol. Cybern.* 82, 247–269. doi: 10.1007/s004220050024

Wehner, M., Truby, R. L., Fitzgerald, D. J., Mosadegh, B., Whitesides, G. M., Lewis, et al. (2016). An integrated design and fabrication strategy for entirely soft, autonomous robots. *Nature* 536, 451–455. doi: 10.1038/nature19100

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

Wilson, D. S., and Sober, E. (1989). Reviving the superorganism. *J. Theor. Biol.* 136, 337–356. doi: 10.1016/S0022-5193(89)80169-9

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Sims. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

HYPOTHESIS AND THEORY
published: 27 April 2022
doi: 10.3389/fnbot.2022.850489

# Resonance as a Design Strategy for AI and Social Robots

*James Derek Lomas[1]*, Albert Lin[2], Suzanne Dikker[3,4], Deborah Forster[2], Maria Luce Lupetti[1], Gijs Huisman[1], Julika Habekost[5], Caiseal Beardow[1], Pankaj Pandey[6], Nashra Ahmad[6], Krishna Miyapuram[6], Tim Mullen[7], Patrick Cooper[8], Willem van der Maden[1] and Emily S. Cross[9,10]*

[1] Department of Human Centered Design, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands, [2] Center for Human Frontiers, Qualcomm Institute, University of California, San Diego, San Diego, CA, United States, [3] Department of Psychology, New York University, New York, NY, United States, [4] Department of Clinical Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [5] The Design Lab, California Institute of Information and Communication Technologies, University of California, San Diego, San Diego, CA, United States, [6] Centre for Cognitive and Brain Sciences, Indian Institute of Technology, Gandhinagar, India, [7] Intheon Labs, San Diego, CA, United States, [8] Department of Physics, Duquesne University, Pittsburgh, PA, United States, [9] Social Robotics, Institute of Neuroscience and Psychology, School of Computing Science, University of Glasgow, Glasgow, United Kingdom, [10] SOBA Lab, School of Psychology, Macquarie University, Sydney, NSW, Australia

Resonance, a powerful and pervasive phenomenon, appears to play a major role in human interactions. This article investigates the relationship between the physical mechanism of resonance and the human experience of resonance, and considers possibilities for enhancing the experience of resonance within human–robot interactions. We first introduce resonance as a widespread cultural and scientific metaphor. Then, we review the nature of "sympathetic resonance" as a physical mechanism. Following this introduction, the remainder of the article is organized in two parts. In part one, we review the role of resonance (including synchronization and rhythmic entrainment) in human cognition and social interactions. Then, in part two, we review resonance-related phenomena in robotics and artificial intelligence (AI). These two reviews serve as ground for the introduction of a design strategy and combinatorial design space for shaping resonant interactions with robots and AI. We conclude by posing hypotheses and research questions for future empirical studies and discuss a range of ethical and aesthetic issues associated with resonance in human–robot interactions.

**Keywords: resonance, entrainment, synchronization, metaphor, design space, social robotics, AI for wellbeing, human-media interaction**

OPEN ACCESS

***Edited by:***
*Adam Safron, Johns Hopkins Medicine, United States*

***Reviewed by:***
*Guillaume Dumas, Université de Montréal, Canada*
*Vicente Raja, Western University, Canada*

****Correspondence:***
*James Derek Lomas*
*j.d.lomas@tudelft.nl*

***Received:*** *07 January 2022*
***Accepted:*** *23 March 2022*
***Published:*** *27 April 2022*

***Citation:***
*Lomas JD, Lin A, Dikker S, Forster D, Lupetti ML, Huisman G, Habekost J, Beardow C, Pandey P, Ahmad N, Miyapuram K, Mullen T, Cooper P, van der Maden W and Cross ES (2022) Resonance as a Design Strategy for AI and Social Robots. Front. Neurorobot. 16:850489. doi: 10.3389/fnbot.2022.850489*

## INTRODUCTION

Resonance is a powerful physical mechanism that manifests in any physical system involving oscillations (Buchanan, 2019). Examples include the electromagnetic resonances that enable wireless communications, the acoustic resonances that give musical instruments their beauty, and the orbital resonances that shaped our solar system. No matter the medium, resonance produces amplification and synchronization effects in oscillatory systems. Details on the varying kinds of resonance are found in **Box 1**.

This article reviews the role of resonance in human systems, in AI and in human–robot interactions (HRI). Given the general appreciation of resonance in human interactions, we argue that designers can make use of the untapped potential of resonance to shape successful and desirable interactions in AI and HRI.

**BOX 1** | A compilation of "resonance" terms from the scientific literature.

The following table outlines the breadth of the concept of resonance across three domains in the social and physical sciences. The examples given for each type of resonance (right column) are not meant to be an exhaustive reference list—instead, our intention is to include a few illustrative examples of each conception of resonance.

| | |
|---|---|
| **Psychology and neuroscience** | |
| Affective resonance | Decety, 2010; Mühlhoff, 2019 |
| Bodily resonance | Bedder et al., 2019 |
| Conceptual resonance | Lee et al., 2007; Howie and Bagnall, 2020 |
| Cognitive resonance | Giorgi, 2017 |
| Embodied resonance | Kirsch et al., 2016b; Gallese and Sinigaglia, 2018 |
| Emotional resonance | Gratch et al., 2013; Schrock et al., 2004; Decety, 2010; Giorgi, 2017 |
| Empathic resonance | Azevedo et al., 2013 |
| Harmonic resonance | Lehar, 2003 |
| Interpersonal resonance | Uithol et al., 2011; Himberg et al., 2018 |
| Intrapersonal resonance | Uithol et al., 2011 |
| Limbic resonance | Lewis et al., 2001 |
| Motor resonance | Cross et al., 2006; Aglioti et al., 2008 |
| Neural resonance | Large and Snyder, 2009; Katz, 1995 |
| Neuroaesthetic resonance | Beardow, 2021 |
| Perceptual resonance | Schütz-Bosbach and Prinz, 2007 |
| Physiological resonance | Engert et al., 2019 |
| Social resonance | Kopp, 2010; Wheatley and Sievers, 2016 |
| **Other social sciences** | |
| Advertising resonance | McQuarrie and Mick, 1992 |
| Aesthetic resonance | Farber, 1994 |
| Brand resonance | Keller, 2010 |
| Carnal resonance | Paasonen, 2011 |
| Consumer resonance | Shang et al., 2017 |
| Cultural resonance | McDonnell et al., 2017 |
| Entrepreneurial resonance | Warren, 2004 |
| Ethical resonance | Prasad, 2019 |
| Frame resonance | Snow and Benford, 1988; Giorgi, 2017 |
| Historical resonance | Ferreira and Vale, 2020 |
| Human resonance | Rosa, 2018 |
| Interaction resonance | Hummels et al., 2003 |
| Institutional resonance | Strydom, 2003 |
| Morphic resonance | Sheldrake, 2011 |
| Narrative resonance | van Werven et al., 2019; Duarte, 2013 |
| Norm resonance | Gutterman, 2015 |
| Political resonance | Cunneen, 2019 |
| Sexual resonance | Baudrillard, 2005 |
| Spiritual resonance | Siegel, 2013 |
| Value resonance | Schemer et al., 2012 |
| **Physics** | |
| *Types of resonance discussed in physics literature include:* | |
| Antiresonance | Rajasekar and Sanjuan, 2016 |
| Autoresonance | Rajasekar and Sanjuan, 2016 |
| Chaotic resonance | Rajasekar and Sanjuan, 2016 |
| Coherence resonance | Rajasekar and Sanjuan, 2016 |
| Ghost resonance | Rajasekar and Sanjuan, 2016 |
| Harmonic resonance | Li et al., 2020 |
| Multiple harmonic resonance | Ludeke, 1942 |
| Parametric resonance | Rajasekar and Sanjuan, 2016 |
| Stochastic resonance | Rajasekar and Sanjuan, 2016 |
| Subharmonic resonance | Ludeke, 1942 |
| Sympathetic resonance | Zhang et al., 2013 |
| Vibrational resonance | Rajasekar and Sanjuan, 2016 |
| *Furthermore, as resonance occurs in any physical system with oscillations, there are medium-specific resonances, including the following examples:* | |
| Acoustic resonance | Ziada and Lafon, 2014 |
| Chemical resonance | Freeman et al., 2014 |
| Electrical resonance | Blanchard, 1941 |
| Friction resonance | Duan et al., 2021 |
| Geometrical resonance | McMillan and Anderson, 1966 |
| Gravitational resonance | Baeßler et al., 2015 |
| Magnetic resonance | Slichter, 2013 |
| Mechanical resonance | Wilfinger et al., 1968 |
| Optical resonance | Oldenburg et al., 1998 |
| Orbital resonance (mean-motion resonance) | Sinclair, 1975; Wang et al., 2021 |
| Plasma resonance | Dahm et al., 1968 |
| Quantum resonance | Moran et al., 2017 |
| Reaction resonance | Yang et al., 2015 |
| Tidal resonance | Garrett, 1972 |
| *Additionally, resonances can emerge from the combinations of basic physical forces, such as those illustrated by the following examples:* | |
| Electromagnetic resonance | Fauché et al., 2017 |
| Nuclear magnetic resonance | Hore, 2015 |
| Plasma-electron resonance | Tonks, 1931 |
| Spin-mechanical resonance | Poshakinskiy and Astakhov, 2019 |
| Magneto-mechanical resonance | Grimes et al., 2002 |
| Electromagnetic acoustic resonance | Hirao and Ogi, 1997 |
| Nuclear acoustic resonance | Sundfors et al., 1983 |
| Spin gravitational resonance | Quach, 2016 |
| Electron spin resonance | Wertz, 2012 |
| Optical spin resonance | Crooker et al., 1997 |
| *Finally, there are emergent resonances that take on a researcher's name, including the following:* | |
| Fano resonance | Lassiter et al., 2010 |
| Feshbach resonance | Tojo et al., 2010 |
| Mie resonance | Roll and Schweiger, 2000 |
| Proudman resonance | Vilibić, 2008 |
| Schumann resonance | Williams, 1992 |

Most of the forms of resonance listed in the PHYSICS category appear to be based upon Helmholtz's (2009) idea of sympathetic resonance. For instance, in a review of magnetic resonance, "the term resonance implies that we are in tune with a natural frequency of the magnetic system" (Slichter, 2013). Yet a recent Royal Society review article (Vincent et al., 2021) makes the following claim: **"the definition of resonance has been generalized [to include] all known processes leading to the enhancement, suppression or optimization of a system's response through the variation/perturbation/modulation of any system property."** This incredibly broad definition of resonance in physics suggests the challenge and need for a coherent understanding of this important concept across physics, neuroscience, the social sciences and design.

## Resonance in Human Interactions: A Metaphor, a Mechanism or Both?

"Resonance" is a commonly-used term that describes the human experience of powerful, connecting and activating interactions (Duarte, 2013). For instance, we can "resonate" with a film or with a new friend. Metaphors related to resonance are also common, such as in the expressions "syncing up," "getting on the same wavelength," or even "feeling good vibes."

Although the term "resonance" is often intended as a metaphor to describe an interaction, in many cases physical resonance may also be a mechanism underlying the interaction. For instance, people metaphorically "resonate with music" but

the brain also physically resonates with music: the actual frequencies of sound and rhythm can be observed in the frequencies of electrical activity in the brain (Coffey et al., 2019, 2021; Kaneshiro et al., 2021; Pandey et al., 2021). Or, consider the common expression "syncing up" to describe a meeting. Even though "syncing" (that is, synchronization) is intended purely as a figure of speech, human communication does link to measurable "inter-brain synchrony" (Dumas et al., 2010; Dikker et al., 2019; Czeszumski et al., 2020; Kingsbury and Hong, 2020; Dumas and Fairhurst, 2021; Moreau and Dumas, 2021). This article aims to create a bridge between the human experience of resonance and resonance as a physical mechanism. In so far as resonance is more than a metaphor—if resonance is also a causal mechanism in human interactions—then this will have implications for measurement and design.

Some skepticism is justified in viewing resonance in human interactions as "real" rather than as just a metaphor. Historically, sympathetic resonance (sympathy: σύμπάθειά or *sumpátheia*) was viewed as the primary mechanism for magical phenomena. For instance, the neoplatonic philosopher Plotinus (205–270) wrote: "But how do magic spells work? By sympathy [sumpatheiai] and by the fact that there is a natural concord [sumphonian] of things that are alike [homoion] and opposition of things that are different." (Lobis, 2015). Even in the modern era, there remains a widespread belief system that positive thinking or "thought vibrations" can bring about positive real-world occurrences through sympathetic resonance (Atkinson, 1906; Hicks and Hicks, 2006; Ehrenreich, 2009). Perhaps as a result of this association with magic, resonance was not always acceptable as a scientific explanation. A recent column in *Nature Physics* notes:

> "…until the very late nineteenth century, scientists were reluctant to use the term 'resonance' in connection with anything except acoustic phenomena, where it originated. Use of the word in other fields…always included some disclaimer that the link was "only by analogy", despite the formal equivalence of the fundamental dynamical equations." (Buchanan, 2019)

Now, the situation has changed: the term "resonance" is abundant in contemporary scientific literature (reviewed in **Box 1**). However, the term is often used ambiguously, where it is unclear whether "resonance" is being treated as a metaphor or as a physical mechanism. This ambiguity is present in the social sciences as well as in physics. Resonance in physics is an increasingly broad concept that refers to a range of phenomena. To bring clarity, we offer a glossary in **Box 2** with proposed definitions for resonance and related terms, such as synchronization, entrainment, reverberation, etc.

## Resonance as a Physical Mechanism

To provide grounding for resonance in human dynamics, this section outlines physical resonance as a causal mechanism in acoustics. Though we focus on sound, it is important to note that resonance operates in all oscillating systems, regardless of medium. This universality results from the fact that resonance is a mathematical property—it is the natural result of the alignment of phases in oscillating systems.

A wine glass offers an excellent example of the physics of resonance. First, if a glass is gently tapped with a spoon, there will be a *reverberating* sound that reflects the natural frequencies of oscillations in the wine glass. These natural frequencies, which are inherent to the structure of the glass, are also known as characteristic frequencies or eigen frequencies ("eigen" is German for "own" or "inherent"). These natural frequencies are also the *resonance frequencies* of the glass: when external, forced oscillations match these natural frequencies, resonance occurs. But, while tapping the glass with a spoon may reveal the resonant frequencies of the glass, the wine glass is not *in resonance* with the spoon.

*Sympathetic resonance* occurs when external, forced oscillations are aligned to a system's own natural oscillations. If a loudspeaker plays the resonant frequencies of a wine glass, the glass will begin to oscillate at much greater amplitudes than if the speaker played other, non-resonant frequencies. Now the glass is in resonance with the speaker. This effect manifests in other common acoustic systems, as well. When one tuning fork is struck near another identical fork, they will both begin to oscillate together, having been coupled together in synchrony *via* the acoustic vibrations. Similarly, two strings tuned to the same note will move one another in synchrony through sympathetic resonance (**Figure 1**).

The relationship between resonance, synchrony and amplification was articulated by German nineteenth century scientist Hermann Helmholtz. His book "On the Sensations of Tone as a Physiological Basis for the Theory of Music" (Helmholtz, 2009; originally published 1863) offers the first scientific exposition of sympathetic resonance in acoustics. His primary illustration of sympathetic resonance involves the resonance between a church bell and its bellringer. If the bellringer provides consistent pulls at a frequency that aligns to the bell's natural rate of swinging, then the swinging will be rapidly amplified. Importantly, the sympathetic resonance occurs when phases of oscillation align: that is, when the downward pull of the bellringer matches up with the downward motion of the bell's swing. The role of synchrony in sympathetic resonance is easier to observe with a slow bell ringer than with the rapid oscillations of a wine glass. Yet, even the sympathetic, synchronized oscillations of a wine glass can be made visible with high frequency camera equipment (Slow Mo Guys, 2021).

Synchrony between systems does not necessarily imply sympathetic resonance. Two systems might be synchronized with each other due to a third system, for instance, or for other non-causal reasons (Hasson and Frith, 2016). Other forms of resonance only occur with powered oscillators (like the clocks and metronomes in **Figure 1**), namely entrainment and synchronization. These terms—which explain phenomena like the synchronization of fireflies or the entrainment of dancers to a musical beat— are defined and discussed in **Box 3**. In this article, we treat these two terms as subsets of sympathetic resonance (by analogy, like squares are subsets of rectangles).

## Resonance as a Metaphor

Having briefly considered the operation of resonance as a physical mechanism, we now wish to bring clarity to the metaphorical use of resonance in science and broader culture.

**BOX 2** | Glossary of terms.

**Resonance**, in this article, is treated as an umbrella term that involves physical resonance (either sympathetic or internal), synchronization, entrainment, memesis and attunement—as well as metaphorical resonance.

**Sympathetic Resonance** occurs when external, forced oscillations are aligned to a system's own natural oscillations and this results in amplification and synchronization. The amplification effect occurs when the natural frequencies of a system align with the frequencies of an external oscillator. Resonance also involves a synchronization or mirroring effect, as the phase and frequency of an external oscillator are reflected in the phase and frequency of the system's response.

**Internal Resonance** involves the activation of the natural frequencies ("eigen frequencies") of a system. For instance, tapping a wine glass results in internal resonances.

**Synchrony** is a broad term that describes the temporal correlation of independent units of action. Correlations can occur in frequency independently from phase; for instance, heart rate synchrony can occur in people who have the same heart rate, even if their hearts do not beat at exactly the same time. Although all meanings of resonance should refer to a causal phenomena, synchrony can occur without a causal relationship (i.e., correlation does not imply causation).

**Synchronization**, in contrast to synchrony, should be understood as a complex dynamic and causal process—not a state. Pikovsky et al. (2001) carefully define synchronization as "an adjustment of rhythms of oscillating objects due to their weak interaction." Further, synchronization requires "self-sustained oscillators," like a powered metronome. Self-sustained oscillators are those that "oscillate with a distinctive waveform at a preferred amplitude that reflects a balance between energy inflow and dissipation." (Strogatz, 2003). Synchronization also requires a sort of "weakness in coupling." Weakness is important because a very strong coupling between systems simply results in immediate complete synchronization. The synchronization of two oscillators does not require that the two have the same phase at the same time; for instance, two clock pendulums can be synchronized but swing in anti-phase. When an external oscillation frequency is nearly aligned to a natural frequency of a powered oscillatory system, the systems will "phase lock" together and synchronize. Synchronization can occur at "a rational fraction of the resonance frequency," like 2:3 or 2:1 (Shim et al., 2007).

**Entrainment** occurs when a consistent rhythmic pulse of one oscillator shifts the frequency of another self-sustained oscillator. For instance, a drummer's beat can entrain the motion of rowers or entrain dancers to a common rhythm. Like synchronization, entrainment requires weakly-coupled and self-sustained (powered) oscillators (Pikovsky et al., 2001). In fact, the two terms are nearly identical; at least one author (Izhikevich, 2007) claims that entrainment is limited to 1:1 synchronization. According to Helfrich et al. (2019), true entrainment requires that "an ongoing oscillator is entrained by a rhythmic input at a slightly different frequency. The entrained oscillation becomes phase-locked and the amplitude increases. After the entraining stream stops the oscillator exhibits a reverberation at the driving frequency for several cycles." Some definitions of entrainment require that an external oscillator unidirectionally influences a powered oscillator (Lakatos et al., 2019) but other definitions allow for mutual entrainment, "whereby two rhythmic processes interact with each other in such a way that they adjust toward and eventually 'lock in' to a common phase and/or periodicity" (Clayton et al., 2005). In this article, entrainment is treated as the mechanism for synchronization (synchronization through entrainment) and both terms are treated as types of resonance (see **Box 3**).

**Reverberation** can be defined by the reverberation time, which is the time required for an oscillation to "fade away" once the external input has stopped. After an impulse, a system's natural or resonance frequencies tend to continue to reverberate, as in a tapped wine glass or echoes in a cathedral.

**Coherence** refers to the statistical similarity between two or more oscillating systems (Wolf, 2010).

**Mimesis**, or imitation, describes the intentional or non-intentional replication of movement patterns. These replications do not need to be synchronous. For instance, a child sticking out their tongue and another child copying them. This can be viewed as a type of resonance enabled by memory.

- Behavioral Mimicry occurs when people behave in similar or identical ways within a short period of time. (Mayo and Gordon, 2020); i.e., "the replication of automated behaviors"
- Imitation can be described as "a short sequence of actions that I see my interaction partner performing and then consequently replicate…imitation is not mere mirroring in the sense that one copies every little part of another's movement. It is rather the replication of the action with regard to the outcome of the action which leads to the acquisition of new skills" (Lorenz et al., 2016).

**Behavioral synchronization** describes behaviors that are synchronous in time, but potentially complementary (e.g., turn-taking) (Chartrand and Van Baaren, 2009).

**Physiological Synchrony** or Biobehavioral synchrony involves the rhythmic and temporal correlation of breath rate, heart rate, hormone production, or interbrain synchrony (Feldman, 2017; Mayo and Gordon, 2020)

**Psychological Attunement** has been defined as "Entrained rhythms [that] constitute a form of dynamic equilibrium in which partners vary their behaviors over time while keeping this variation within desired limits" (Sadler et al., 2009).

**The Vibe** (e.g., "good vibes" or "vibing with") is a pervasive cultural construct used to describe how people perceive the shared affective experience and aesthetic expectations of a group, a place, a product, a brand, a robot, etc. The vibe is different from a person's individual affective reaction, as it describes the aspects of conscious experience that are perceived to be shared between people (Witek, 2019). Hypothetically, the vibe emerges from interpersonal resonance effects.

**Harmony** is an ancient concept (Lomas et al., 2022) that has an intrinsic relationship with resonance: the harmonic tunings of stringed instruments maximize acoustic resonance between the tuned strings. For example, musical notes with consonant intervals (such as the 2:3 ratio of a musical fifth) will share common acoustic harmonics, while dissonant intervals do not. These shared harmonics produce physical resonances between tuned strings—and perhaps resonances between neural oscillators, as well.

A recent review of the word "resonance" in the language of scientific literature (Ruthven, 2021) reveals that resonance typically serves as an implicit metaphor to indicate 1. agreement (e.g., new evidence can resonate with an existing theory), 2. arousal (e.g., a film that resonates is engaging and moving) or 3. action (e.g., the resonance of a speech can motivate people to take action). But, despite a vast number of scientific articles that use resonance as a term, it is only very rarely defined. The lack of definition suggests that "resonance," as a term, is easily and broadly understood *intuitively* as a metaphor.

Metaphors are useful when they enable concrete, familiar experiences to communicate abstract, conceptual meanings (Lakoff and Johnson, 1980; Yang, 2014). A metaphor involves the pairing or alignment of concepts between a concrete source and a more abstract target; this coupling of concepts produces mappings that allow multiple concepts to be integrated together

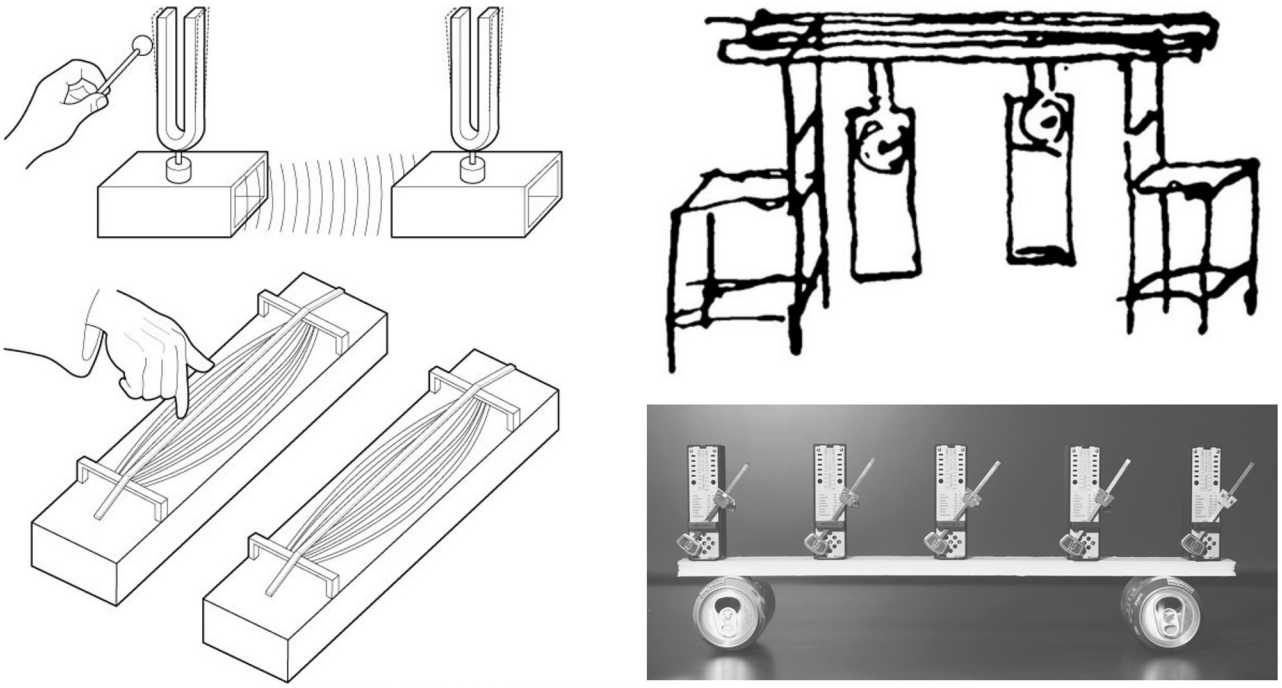


**FIGURE 1 | (Left)** Resonance between mutually attuned tuning forks and strings involves synchronization and amplification. **(Right)** Sketch by Christiaan Huygens (b. 1629), who discovered "the sympathy of two clocks." When two clocks are placed on a common beam, their two pendulums will eventually synchronize. Right bottom: a set of weakly coupled powered metronomes (self-sustained oscillators) will eventually synchronize. Photo courtesy: Harvard Natural Sciences Lecture Demonstrations.

into an emergent space of meaning (Holyoak and Stamenković, 2018). An example set of metaphorical mappings use the metaphor "Love is a Journey": for instance, in a journey (the source) there are travelers while in love (the target) there are lovers. In **Table 1**, we provide a similar set of explicit mappings between the metaphor of acoustic resonance and resonant human interactions.

With this introduction to resonance established, we are now posed to explore the alignment between the metaphorical experience of resonance (as in a film that resonates) and the physical phenomena of resonance itself. The next section of the article considers the research basis for understanding resonance in human interactions.

## PART 1: RESONANCE IN HUMAN INTERACTIONS

A common example of physical resonance in human interactions can be found on nearly any playground. When pushing someone on a swing, the pusher needs to coordinate the timing of their pushes to the swing's natural back-and-forth oscillation (determined primarily by the length of the swing). Does pushing at a faster rate help? No: if the pusher simply pushed more times per second, most of the pushes would do nothing because they would not line up with the movement of the swing. When a pusher aligns their timing to the natural frequency of the swing, they amplify the effects of their effort: many small, well-timed pushes are enough to get the swinger high into the air.

Beyond this simple example, where else might resonance occur in human interactions? To scope our search, we assume that sympathetic resonance can only occur when external oscillations and natural oscillations align. Therefore, physical resonances in human systems should only be present during human activities that have a natural frequency or rhythm of oscillation.

Consider an everyday rhythmic human activity: walking. Researchers have used accelerometers to determine the dominant and natural up-and-down frequency of walking. The typical frequency of naturalistic walking is about 2 Hz, or two steps per second (MacDougall and Moore, 2005). This natural oscillatory frequency can vary—some people walk faster or slower than others. However, across a diverse set of participants, the researchers found that the tempo of walking was not dependent upon height, weight, or other physical factors. In fact, the researchers suggest that the 2 Hz natural tempo is the result of genetically encoded "central pattern generators" in the spinal cord, as these are the basis of the tempo of locomotion in other animals (Guertin, 2013).

This is not just trivia: structural engineers need to take into account this 2 Hz human walking pace in every footbridge that is built. Famously, a 2 Hz resonance frequency caused the UK's Millenium Bridge to dramatically sway side-to-side when it was loaded with pedestrians (Dallard et al., 2001; Strogatz et al., 2005). On opening day, the bridge became so crowded that most people could not easily walk forward—instead people were so packed-in that they had to walk in a sort of side-to-side waddle. Unfortunately, the bridge had a natural side-to-side 2 Hz

**BOX 3** | Clarifying the relationship between resonance, entrainment and synchronization.

This article treats resonance as an umbrella concept that includes both metaphorical resonance and resonance as a physical mechanism. As resonance has an expansive meaning in physics (**Box 1**), we first distinguish between typical resonance (which involves the natural frequencies of a system) and atypical resonance (which does not involve natural frequencies). Then, typical resonance includes *internal resonance* (the activation of reverberating natural frequencies) and *sympathetic resonance* (alignment between the frequencies of external oscillations and the natural frequencies of a system). Sympathetic resonance can be further divided into *passive resonance* and *active resonance*. Passive resonance occurs with unpowered oscillators (like a wine glass) while active resonance occurs with powered or *self-sustained oscillators* (like battery-powered metronomes). Active resonance includes the phenomenon of *entrainment*, which occurs when a self-sustained oscillator synchronizes its phase and frequency to a weakly coupled external oscillation.

The natural or resonance frequency of an oscillator can be considered its *preferred frequency*. Powered oscillators, like metronomes, also have a *preferred amplitude* of oscillation—this is not the case for passive oscillators, like wine glasses. When external frequencies align with the preferred frequencies of a wine glass, the most noticeable aspect is the amplification of the amplitude of the oscillation. In contrast, when external oscillations align with the natural frequencies of a powered oscillator, the most noticeable aspect is the synchronization. However, the synchronization of frequency and phase also occurs between a wine glass and an external speaker while amplification also occurs with synchronized metronomes. For this reason, we describe entrainment and synchronization as types of sympathetic resonance; namely, the type involving a self-sustained oscillator.

The scientific relationship between entrainment and resonance is often a point of confusion due to the lack of clear definitions (Helfrich et al., 2019). Our view diverges from other descriptions of resonance that are limited to passive, unpowered systems (Pikovsky et al., 2001; Guevara Erra et al., 2017; Lakatos et al., 2019). Our view is that the concept of resonance can easily accommodate active forms as well as passive forms, as both involve preferred frequencies of oscillation (natural frequencies), synchronization effects and amplification effects. Given the ubiquity of resonance in oscillatory systems—and its already expansive definition in physics (**Box 1**)—why should resonance only refer to unpowered systems and thus exclude dissipative systems, like the brain? Rather than treating "resonance" in interpersonal interactions as a complete misnomer, we make the case that it is appropriate and physically accurate to say that we *resonate* with people, films or other media. We hope that this view opens the door to a more comprehensible and coherent scientific study of resonances in human interactions.

**Types of Resonance**

- Metaphorical Resonance
- Physical Resonance
  - Atypical Resonance: Does not involve natural frequencies (see Vincent et al., 2021)
  - Typical Resonance: Involves natural frequencies
    - Internal Resonance (involves the activation and reverberation of natural frequencies within a stimulated system, like the reverberations of a tapped wine glass)
    - Sympathetic Resonance (involves the alignment of external frequencies with the natural frequencies of a system—when the forced frequencies match the natural frequencies)
      - Passive Resonance: unpowered and externally sustained; like a wine glass vibrating in synchrony with external oscillations.
      - Active Resonance: powered and self-sustained, like a metronome synchronizing with external oscillations. This encompasses different types of synchronizatn:
        - Complete Synchronization (due to strong coupling)
        - Entrainment (Phase Synchronization)
          - In-phase
          - Anti-phase
          - Phase shifted
        - Frequency Synchronization
        - Envelope Synchronization
        - Partial and Asynchronous Synchronization (e.g., mimesis)

resonant frequency. As the bridge started to sway back and forth, this led to the synchronization of the waddling motion of the thousands of pedestrians. Without deliberate coordination, people stepped left and stepped right in synchrony with each other, entrained to the swaying motion of the bridge. As a result, the 2 Hz side-to-side oscillation of the bridge was further amplified to dangerous levels.

Given that 2 Hz is a *natural* frequency of human movement, the theory of resonance predicts that 2 Hz should also be a resonance frequency. That is, external rhythmic inputs at about 2 Hz should cause a synchronization and amplification of human movement. Is 2 Hz actually a resonance frequency of movement? One way to test this prediction is to consider the popularity of music at different beats per minute (BPM), where 120 BPM would be equivalent to 2 Hz. **Figure 2** presents a histogram showing the relative distribution of BPM in the Top 50 songs for each year of the past decade, worldwide. This reveals a distinct peak at 120 beats per minute.

Is this a resonance effect? The graph in **Figure 2** resembles a resonance curve: given a range of input frequencies, there is a selective amplification at the same frequencies as the natural frequencies of the stimulated system (i.e., the 2 Hz natural pace of movement in humans). However, the peak may arise for a variety of reasons, from the musicians' recognition of the popularity of 120 BPM to the listeners' increased familiarity with 120 BPM songs. Most importantly, there are many more songs released at 120 BPM. So, even though it is not entirely appropriate to refer to the graph as a resonance curve, it may still be a function of resonance effects (e.g., a greater likelihood of rhythmic entrainment to walking frequencies at 120 BPM).

In this article, we propose that physical resonance can be distinguished from purely metaphorical resonance when the mathematics of resonance (as used to characterize other physical systems) can be used to model human interactions. Synchronization, amplification and signal alignment are the mathematical hallmarks of resonance—therefore, physical resonance in human systems should occur when the oscillations of external signals match natural human oscillations and this results in synchronization and increased energy (amplification). This viewpoint treats the physics of human resonance in a similar fashion as other physical systems yet it leaves room for future research to further clarify the relationships involved.

**TABLE 1 |** A hypothetical conceptual mapping of acoustic resonance and human resonance.

| Acoustic resonance (source) | Human resonance (target) |
|---|---|
| A string tuned to another string can respond sympathetically | A person attuned to another person can respond sympathetically |
| A string's sympathetic response to another string results in synchrony | A person's sympathetic response to another person results in synchrony |
| A string mirrors the vibrations of another string | A person mirrors "the vibe" of another person (see **Box 2**) |
| A string vibrates to the frequency of another string, if tuned | A person responds to the expression of another person, if attuned |
| The attunement of strings is based on a common or an aligned set of oscillations | The attunement of people is based on a common background or an aligned set of experiences |
| Tuning strings enhances the resonance between strings | Attuning people (e.g., with common experiences) enhance resonance |
| Resonance results in greater amplitude of sound | Resonance results in greater excitement in people |
| A string will only selectively resonate to particular frequencies, based on its own natural oscillations | A person will only selectively resonate to particular [people, films, books, etc], based on their own natural propensities |

## Rhythmic Human Interactions

The previous example shows how rhythmic human activity can be investigated as a context for physical resonance phenomena. Rhythmic human interactions clearly occur in artistic domains, such as music-making (Clayton et al., 2005), dance (Larsson et al., 2019) and various kinds of cultural rituals like chanting (Gelfand et al., 2020). Rhythmic interactions are also common in everyday life, as in the case of walking, conversational turn-taking (McGarva and Warner, 2003; Wilson and Wilson, 2005; Lee et al., 2010), patterns of eye contact (Wohltjen and Wheatley, 2021) or with interactions like handshakes (Melnyk and Hénaff, 2019). More intimate rhythmic interactions occur during human sexual behavior (Safron, 2016). Leading up to the moment of birth, midwives often advise expectant mothers to push in phase synchrony with their own rhythmic uterine contractions (Hanson, 2009). Researchers have also observed that the earliest interactions between parent and child are strongly rhythmic (Stern et al., 1985). Babies cry in a rhythmic manner

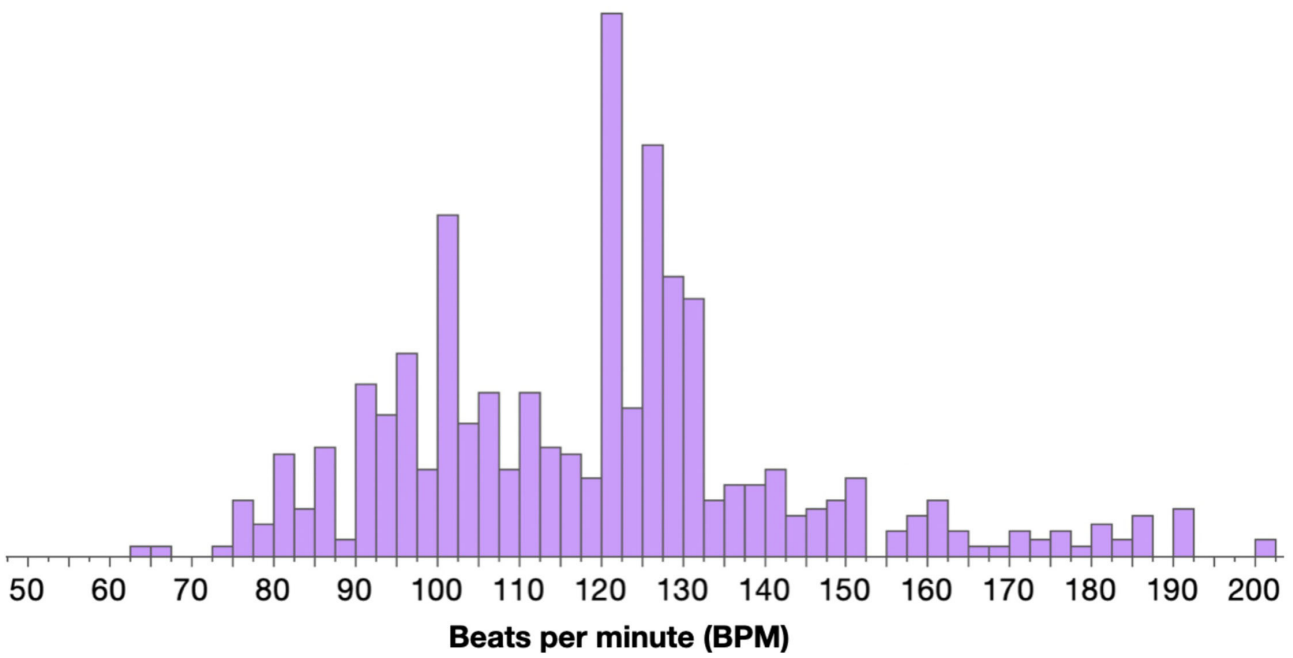



**FIGURE 2 |** The above histogram is compiled from a Kaggle dataset containing the "Top 50" Billboard songs from 2010 to 2019. It appears to be a resonance curve showing maximum excitement at a preferred frequency of oscillation. However, note the sharp dropoff from 120 BPM to 118 BPM—this is not expected in a resonance curve, as the frequencies close to resonance tend to resonate strongly as well. There may simply be very few songs released with this BPM. But, while this graph might not be a resonance curve, it may be a result of meaningful resonance effects in the brain—an increase in amplitude due to the alignment of external frequencies with natural frequencies https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year.

and caregivers soothe them with synchronized motions (Trehub and Trainor, 1998). Synchrony in caregiving appears to literally "tune" the human social brain (Yaniv et al., 2021).

> "This resonance or echoing of affect, feelings, and emotions that takes place in the reciprocal interaction between infants and their caretakers is a necessary element for the development of empathy and advanced social cognition." (Decety and Meyer, 2008)

## Neurobiological Rhythms and Human Nature

What makes human behavior so rhythmic? Human rhythms are believed to emerge from a broad range of biological oscillators that are present across the brain and the rest of the body (Varga and Heck, 2017). For instance, rhythmic central pattern generators in the spinal cord not only drive locomotion (Ijspeert, 2008; Guertin, 2013), but also drive heart rhythms (Bucher et al., 2015) and breath rhythms (Molkov et al., 2014). Furthermore, body and brain oscillations appear to integrate together in a hierarchical architecture (Klimesch, 2018).

Rhythmic—and resonant—phenomena are found in the brain at multiple levels, from neurons to circuits to brain waves (Buzsaki, 2006). Individual neurons have natural intrinsic oscillatory periods (Lampl and Yarom, 1997); neurons can be *tuned* to respond to different input frequencies through resonance, like "strings on a violin" (Das et al., 2017). At the level of neural circuits, the *reverberation* of recurrent activity in neural ensembles plays a key role in memory (Wang, 2001; Tegnér et al., 2002; Han et al., 2008), as originally predicted by neuroscientist D.O. Hebb in 1949 (Hebb, 2005). Finally, large-scale electrical oscillations in the brain, or brainwaves, demonstrate clear resonance effects (Herrmann, 2001) that are observable through electroencephalography (EEG). Further to this, neurobiological processes associated with adaptive learning (Grossberg, 2013, 2017), perceptual learning (Raja, 2020), and ecological cognitive architecture more generally (Raja, 2018), have all been theorized as forms of physical resonance.

Neurons are, technically speaking, non-linear oscillators (Izhikevich, 2007; Stiefel and Ermentrout, 2016)—therefore, it is not so surprising that large collections of neurons exhibit both internal resonance effects (one part of the brain resonating to another) and external resonance effects (the brain resonating to environmental phenomena). Many scientists believe that physical resonance in the brain plays a major role in music perception, such as the physicist and neuroscientist Ed Large, who claims:

> "The brain does not 'solve' problems of missing fundamentals, it does not 'compute' keys of melodic sequences, and it does not "infer" meters of rhythmic input. Rather, it resonates to music… certain aspects of this process can be described with concepts that are already well-developed in neurodynamics, including oscillation of neural populations, rhythmic bursting, and neural synchrony." (Large, 2010)

Some researchers have recently proposed a unified account of rhythmic synchronization and entrainment in the brain (Lakatos et al., 2019); other researchers have proposed a unified account of the biological, neurological and physical mechanisms involved in the "rhythmic entrainment of biological systems" (Damm et al., 2020). Rhythmic entrainment has been found to govern patterns of interaction at a social, population, and even species level—where, in the latter case, the entrainment of natural oscillations can be observed at the scale of economies and ecosystems (Greenfield et al., 2021). In short, it would appear that resonance effects can operate all the way up and all the way down: from neurons to economies.

## Entrainment and Rhythmic Synchronization

Human interactions can naturally synchronize through the process of entrainment (**Boxes 2, 3**), which is akin to the natural synchronization of metronomes (**Figure 1**). Social neuroscientist Ruth Feldman (2012, 2017) argues that biobehavioral synchrony (in behavior, heart rate, endocrine production and brainwaves) serves as a key principle underlying parental love, romantic love, friendship and human attachments. Indeed, when loving human partners interact, their rhythmic communication produces measurable physical synchronization in behavior (Grafsgaard et al., 2018), in heart rate (Prochazkova et al., 2022) and in the brain (Kinreich et al., 2017).

Some argue that the ability to synchronize to a beat is one of the core skills associated with human social behavior. Kirschner and Tomasello (2009) found that children 2–4 years old could adjust their natural drumming tempo to match another beat—but that their accuracy in synchronizing was significantly higher when they drummed with a human partner (as opposed to drumming along with a machine or drumming along with a drum sound produced by a speaker). The authors argue that "drumming together with a social partner creates a shared representation of the joint action task and/or elicits a specific human motivation to synchronize movements during joint rhythmic activity."

Humans are typically much less able to synchronize to rhythms of visual flashes than to rhythms of auditory tones (see review by Repp and Su, 2013). But, rhythmic entrainment and synchronization is not specific to music or auditory experiences. With certain forms of visual stimuli (i.e., bouncing balls), visual synchronization becomes nearly as accurate as auditory synchronization (Iversen et al., 2015). Researchers have also found that deaf individuals exhibit enhanced synchronization to visual rhythms, suggesting that the ability to attune to rhythms is at least partially based on experience and not just a result of biological coupling between the auditory and motor system (Iversen et al., 2015). Researchers have found that humans can synchronize to tactile pulses on their back with higher accuracy when feeling the vibrations played over their entire back rather than at just a small portion; similarly, rhythms that engage multiple sensory modalities also produce more accurate synchronization (Ammirante et al., 2016). This suggests that overall sensory immersion and attentional engagement affects the propensity to synchronize with rhythms.

Synchronization helps support coordinated actions between individuals. Meta analyses (Morgan et al., 2017) have shown

that behavioral synchrony in groups increases prosocial behavior, increases perceived social bonding, and generally *feels* good (as measured as increases in positive affect). Why might behavioral synchrony feel good? Cracco et al. (2021) claim that "synchrony is aesthetically pleasing and a signal of group cohesion, as stimuli that are processed more fluently are known to produce a hedonic response."

Rhythmic synchronization is very rare in the animal kingdom, at a social level, apart from special examples (see extended discussion in **Box 4**). The human capacity for rhythmic synchronization may have coevolved in human cultures because it enhanced social bonding between sexual partners, between parents and children, and within larger social groups. Savage et al. (2021) state:

> "'Neural resonance' (synchronous brain activity across individuals) facilitates social bonding through shared experience, joint intentionality, and 'self-other merging'. Through the production of oxytocin and endogenous opioids, neural resonance also facilitates prosociality."

## Hyperscanning and Inter-brain Synchrony

The scientific understanding of rhythmic entrainment and neural resonance is a fast-moving area of neuroscience that is being propelled by new hyperscanning methods that scan the brains of multiple interacting participants simultaneously. Interpersonal neural synchrony at the group and dyadic level has been shown to be associated with a number of predictors, including shared stimulus features, joint actions, personality traits, social intentionality, relationship quality, and cooperation (see, e.g., Czeszumski et al., 2020 for review).

For example, recent work from co-authors of this article investigated the relationship between inter-brain synchrony and group dynamics and found that EEG inter-brain synchrony predicted collective performance among teams better than self-report (Reinero et al., 2021). In another line of work, group-based inter-brain coherence predicted class engagement and social dynamics in groups of high school students during their real-world lessons (Dikker et al., 2017; Bevilacqua et al., 2019). Social closeness with the teacher also correlated with brain-to-brain synchrony—that is, enhanced synchrony was found with students who reported greater engagement with the teacher. Finally, and perhaps most directly related to the concept of resonance: Brainwaves of students who engaged in face-to-face interactions *before* class were more synchronized *during* class, even if students were no longer interacting. This finding raises interesting questions about the role of resonance in the directionality of the relationship between human face-to-face interaction, inter-brain synchrony, and social connectedness.

## The Downside of Being in Sync: Chained to the Rhythm?

Humans may be predisposed to synchronize with each other, but this does not always lead to positive outcomes. Synchronization also has some important tradeoffs; Gelfand et al. (2020) claim that synchrony can produce conformity, destructive obedience, groupthink, antisocial aggression and also impair group creativity. They point to findings (Wiltermuth, 2012a,b) that people who have been randomly assigned to a synchronous activity are more likely to comply with an anti-social order (e.g., irritating a stranger) and to follow a morally compromised command (in the study, participants were asked to grind up live bugs). Synchrony also increases the likelihood that people will engage in conformity, like copying majority opinions rather than following their personal preferences (Dong et al., 2015). Further, sometimes synchrony is simply "situationally inappropriate;" in a study of a complex verbal coordination, groups that were randomly assigned to a synchronization task performed worse, reported higher levels of conflict and reduced group cohesion (Wood et al., 2018).

Gelfand et al. (2020) randomly assigned participants to march synchronously around a college campus or at their own pace. The participants who synchronized showed reduced creativity when writing stories. They also found that synchronous marching discouraged the development and sharing of minority perspectives during decision-making. They relate this finding about synchrony to the need to balance "tightness" and "looseness" in culture.

The ability to flexibly move in and out of synchrony appears to be critical to adaptive flexibility. Mayo and Gordon (2020) claim that "two tendencies exist simultaneously, one to synchronize with others and another to move out of synchrony and act independently. We suggest that an adaptive interpersonal system is a flexible one, able to continuously adjust itself to the social context."

Savage et al. (2021) point out the key difference between rhythmic integration and pure synchronization: rhythm is predictable but also flexible to accommodate diverse individual contributions. This is because rhythm involves two essential components: 1. equally timed beats (isochronicity) and 2. a hierarchical structure (meter). "While synchronization solely to the beat (e.g., in marching or unison chanting) allows large groups to integrate, it tends to submerge individual contributions. Meter solves this problem by allowing many individuals to contribute, out of phase, to the same integrated rhythm." Social rhythms (of speech, music, dance, etc.) can thus support diversely coordinated actions within a loosely unified structure.

## Origins of Empathy: Sympathetic Resonance

Sympathetic resonance—including synchronization and rhythmic entrainment—appears to have been a key factor in human evolution (Savage et al., 2021; Lin and Lomas, 2022). Resonance relates in a fundamental way to the human capacity *to feel what another person feels*, which is often called empathy. But, before the term "empathy" was coined in the twentieth century, the ability to feel what others feel was referred to as "sympathy"—as in sympathetic resonance. The eighteenth century philosopher Adam Smith wrote his first book, "The Theory of Moral Sentiments" (Smith, 1759), with the general thesis that "sympathy" accounts for a large portion of moral behavior. Specifically, he explained that people like to help other

**BOX 4** | Resonance in non-human animals.

Despite the relative utility of resonant phenomena in humans (such as social synchronization and entrainment), it is rare in the animal kingdom. There are examples of animals synchronizing with other members of their species: for instance, chirping insects, croaking frogs, claw-waving crabs, and flashing bioluminescent animals, like fireflies (Wilson and Cook, 2016). However, these examples seem to involve somewhat involuntary neurological connections in fairly simple animals. Why is entrainment not more common? While this is not well-understood at present, Wilson and Cook (2016) provide three criteria for rhythmic entrainment: 1. an animal needs to have the mechanical ability to move with the beat (i.e., the tempo should be similar to an animal's natural tempo of movement), 2. the animal must be able to extract the beat from the sensory environment and pay attention to it, and, crucially, 3. the animal must have the *motivation* to voluntarily move in union.

Until very recently, it was believed that only human beings could synchronize to an external rhythm, like a musical beat. It was only with the advent of YouTube that researchers first discovered Snowball the Dancing Cockatoo (Patel et al., 2009; Patel and Iversen, 2014) and 33 other examples of animals that appeared to show entrainment to music (Schachner et al., 2009). These examples spanned 14 different bird species—and an Asian Elephant. This led to the belief that only species that had previously evolved the capacity for vocal mimicy could entrain to a beat. Schachner et al. (2009) noted that, despite a large number of "dancing" dog videos, none demonstrated the ability to synchronize with music (even though some dogs had been trained for years to compete in dance competitions).

Some animals can be reliably trained to synchronize to a rhythm. In 2011, researchers demonstrated that Budgerigars, a parrot-like bird, could learn to produce rhythmic beak tapping patterns that synchronized to an audio-visual metronome (Hasegawa et al., 2011). Then researchers managed to train Ronan, a Sea Lion (Cook et al., 2013), to entrain to a beat—this was surprising because Sea Lions are not vocal learners.

Over the past decade, there has been much investigation of the capacity for non-human primates to entrain to a musical beat. Sounds can induce spontaneous rhythmic swaying in chimpanzees (Hattori and Tomonaga, 2020)—however, this swaying effect occurs in response to randomized rhythms and when sounds are rhythmic (Bertolo et al., 2021). Monkeys have been trained to tap in response to an auditory or visual metronome, however, their movements are always reactive: they always tap following the stimuli (although much faster than they can in a single reaction time experiment; see Wilson and Cook, 2016). In contrast, when humans entrain to a similar metronome, they typically tap slightly *before* each stimulus in the beat. In just one case, researchers have trained monkeys to make predictive, synchronized eye movements to a visual metronome—however, the monkeys had to be rewarded for each trial (Takeya et al., 2017). Based on this evidence, the authors suggest that monkeys and other animals may have the capacity for "predictive and tempo-flexible synchronization to a beat" but might not be "intrinsically motivated" to synchronize!

In summary, it is surprising that so few animals—neither dogs nor monkeys—are predisposed to entrain to a beat. After all, even animal *neurons* have the capability to entrain to periodic rhythms. Why, then, are animals generally so unable—or unwilling— to entrain to a beat? One possibility: consider that the heart is entrained to rhythms produced by central pattern generators in the spinal cord; clearly, animals need to protect their heartbeat from becoming entrained to external stimuli. It may be that, even in the simplest of animals, there is a need to evolve *defense mechanisms* that can protect against unwanted resonance effects. Part of the human capacity for rhythmic entrainment may result from the ability to "let one's guard down" in order to open up to certain kinds of external rhythmic entrainment with other people. This would suggest that humans only resonate to external stimuli when they feel safe to do so; after all, stress may make it difficult to dance or to be moved by music. This also suggests that animals may be able to resonate, if they could be emotionally or biochemically prepared to do so. This opens up possibilities for animal-robot and animal-AI interactions that can be explored in the future.

people because they sympathetically feel good when other people feel good and sympathetically feel bad when other people feel bad (Schliesser, 2015).

Later, the nineteenth century German psychologist Theodore Lipps used the German term *Einfühlung* to describe how people "feel into" the states of other people and even art pieces. By using an inner imitation or simulation, people seem to be able to fuse with artworks or persons through a process of "Psychische Resonanz" (Lipps, 1891). For instance, watching a tightrope walker produces a resonance with internal associated feelings like vertigo. The representation of the performer in one's own mind allows one to feel how oneself would feel in the same situation. The psychologist Edward Titchener reviewed Lipps' work in 1909 (Titchener, 1909) and, rather than using the German *Einfühlung*, he coined the new English word *Empathy* (Schliesser, 2015).

## Empathy, Motor Resonance and "Mirror Neurons"

Empathy is viewed as a critical component of human social interactions. However, it is extremely challenging to pin down. While there is an enormous amount of scientific work on empathy, there is still considerable debate about its definition (Hall and Schwartz, 2019). Is empathy a singular capability or does it result from a "laundry list" of characteristics? Psychologists generally accept the division between *cognitive empathy* and *affective empathy*. Cognitive empathy refers to the ability to recognize and understand another person's mental state (cognitive processes captured by what is referred to as "theory of mind" or mentalizing), while affective empathy refers "the ability to vicariously experience the emotional experience of others" (Reniers et al., 2011). Furthermore, psychologists will often draw another distinction between empathy (which involves the ability to distinguish the experience of another person's emotion from one's own emotional state) and *emotional contagion*. Emotional contagion involves the direct propagation of emotional states; unlike empathy, this effect is common in non-humans, like mice (Hernandez-Lallement et al., 2020).

Regardless of definition, the capacity for empathy (or, at least, affective empathy) is typically conceptualized as emerging from *motor resonance*. Motor resonance describes how the spatial-temporal activations of an observer's brain mirrors the brain activations of another person as they perform some set of actions. That is, when observing the physical behavior of another person, the brain regions related to this behavior activate in both the observer and the person enacting the behavior, creating a sort of spatial-temporal synchrony between observers and actors. Thus, motor resonance is a type of physical resonance that provides a mechanism for sharing conscious experiences between people. For instance:

> "…the coupling between action and perception, also named "motor resonance" [involves] the automatic activation, during actions perception, of the perceiver's motor system. During action

> observation the two motor brains "resonate" because they share a similar motor repertoire". (Sciutti and Sandini, 2017)

Several theories of empathy describe motor resonance as the mechanism underpinning the *mirroring processes* of emotions and actions, where mirroring processes are defined as automatic processes for internal imitation (Iacoboni, 2009) or embodied simulation (Gallese, 2009). The simulation theory of empathy (e.g., Preston and De Waal, 2002) suggests that people can feel what other people are feeling because observing another person's behaviors will coactivate or call upon neural representations of one's own bodily experience (Hurley, 2008).

While stereotypical expressions may produce meaning symbolically (e.g., a smile is symbolically associated with joy or a frown with sadness), human emotional expression is far more dynamic, expressive, and context dependent. The spatial-temporal dynamics of, say, a rapidly lifted eyebrow can express inner emotional states with great specificity. Observing the eyebrow rapidly lifting will engage our own motor cortex to activate in a similar time scale and specifically in the areas expected for the muscles involved. These spatial-temporal activations appear capable of automatically triggering associated emotional states (Wood et al., 2016). That is, whatever feelings might have been associated with rapid eyebrow lifting in the past, either in the self or in others, will now be primed. In this manner, interpersonal motor resonance appears to support "mind reading" (Agnew et al., 2007) and the sharing of conscious experiences (Lin and Lomas, 2022). Because our brains reflect or mirror each other, through resonance, we can sympathetically experience the feelings associated with other's actions, in part by knowing "how we would feel if we were acting that way." And it is not just observation: even listening to descriptions of actions can trigger motor resonance (Zwaan and Taylor, 2006).

Researchers continue to debate the origins of mirroring processes, but they appear to result from simple bidirectional associations between perception and motor responses that are learned over time (Keysers and Gazzola, 2009, 2014; Hanuschkin et al., 2013). Simple correlations of associated actions and observations seem to produce "action perception circuits" that serve as the neural mechanism for mirroring processes (Pulvermüller, 2018).

For a clear example, fMRI results show that when people watch others perform actions with their hands, mouths or feet, there are activations in their own premotor cortex—activations that are also triggered when performing those actions themselves. Furthermore, these action activations occur "following a somatotopic pattern which resembles the classical motor cortex homunculus." (Buccino et al., 2004a) For instance, if person A watches person B kick a ball, the "leg part" of the premotor cortex will show a similar pattern of activation in person B (the kicker) and in person A (the observer).

## The "Like Me" Hypothesis

If we consider the sympathetic resonance of two tuned strings, the strings have in common their natural frequencies of oscillation. A similar sympathetic resonance occurs when we see another person smile; this can trigger similar action representations in our brain and activate associated emotional states. Following the metaphor of two tuned strings, the "like me" hypothesis predicts that the degree of motor resonance between an actor and observer will correlate with the degree of *similarity* between the actor and the observer.

Buccino et al. (2004b) investigated human motor resonance in response to dogs, monkeys and people. They found that "Actions belonging to the motor repertoire of the observer (e.g., biting and speech reading) are mapped on the observer's motor system. Actions that do not belong to this repertoire (e.g., barking) are essentially recognized based on their visual properties." In other words, actions that are not "like me" may be recognized but they do not resonate.

If resonance is enhanced when observing actors similar to the observer, is it also impaired when there is a lack of similarity? Researchers have found that when subjects observe people of a different ethnicity, there is significantly less motor resonance than when watching members of the same ethnicity (Gutsell and Inzlicht, 2010; Azevedo et al., 2013). This unfortunate effect is predicted by resonance theory: less similarity, less resonance.

The resonance between actions is dependent upon a person's ability to do those actions. The "like me" hypothesis predicts that persons who are highly trained in a particular skill should be able to resonate with another person trained in the skill, at least to the extent that their action-observation circuits are mutually developed. Work with expert dancers using fMRI (Calvo-Merino et al., 2005; Cross et al., 2006), EEG (Orgs et al., 2008), and facial EMG (Kirsch et al., 2016a) provides evidence in support of this idea that shared learning experiences and shared skills will increase motor resonance.

Evidence against the Like-Me hypothesis comes from the fact that similarity does not always enhance activation intensity. A series of fMRI experiments showed that mirroring processes (also known as the "Action Observation Network" or AON) are more strongly engaged during the observation of *robot-like* motions, both when the motions were performed by actual robots and when people act in a jerky, robot-like manner (Cross et al., 2012). It appears that the relationship between familiarity and neural resonance is not entirely linear because—in part—the perception of novelty (Knight and Nakada, 1998) also amplifies the brain's response to actions (Gardner et al., 2017).

The brain's sensitivity to novelty may help explain why motor resonance is exceptionally amplified when expert dancers observe other expert dancers perform (Cross and Ramsey, 2021). Experts not only have deep familiarity with the movements but they will also have an expert sensitivity to the many small novelties within the expert's individual execution. Familiarity and novelty—though seemingly opposite—both contribute strongly to an aesthetic experience (Hekkert et al., 2003). This may account for why aesthetically valued actions influence motor resonance. Researchers have found that the intensity of activations in the AON (a brain marker of motor resonance) correlates with aesthetic ratings of the observed dances (Cross et al., 2011). Calvo-Merin et al. (2008) also examined the neural response to dance movements and noted that, of five aesthetic dimensions (like-dislike, simple-complex, dull-interesting, tense-relaxed and

weak-powerful), only shifts in liking-disliking correlated with the brain response in the AON.

Thus, several factors can enhance motor resonance, including inter-subject similarity. Generally speaking, motor resonance appears to be enhanced by the overall motivational relevance of the subject-observer interaction; the similarity of the subject, the novelty of the interaction, the aesthetic quality of the interaction, and when the subject is viewed as desirable, powerful or sharing a common goal (Greenberg, 2019). This also implies some possibilities for breakdown and pathologies in psychological relationships due to *misattunement* in interpersonal resonance (Bolis et al., 2017).

## Human Resonance, Exemplified by the Actor Will Smith

To conclude this section on the role of resonance in human interactions, consider this quote by the actor Will Smith explaining his growth as an actor during his performance as Richard Williams, the father of Venus and Serena Williams. The quote illustrates the connection between the body, the communication of emotional depth and how resonance is generated during aesthetic experiences (described here as "vibrations"). It also shows the vast distance roboticists must travel to approach the capabilities of human actors in producing effective sympathetic resonance (for resonance in acting, also see Bogart, 2021).

> "At the core, acting is what you can comprehend emotionally. And when you comprehend it emotionally, do you understand it enough to feel it and create interesting behavior around it? So something like Richard Williams's walk: Now, you can mimic someone's walk and look authentic. It's a completely different thing when you know why the person is hunching over vs. the stand-up-comedian version of it just mimicking it. Understanding that was the leap that happened: When you know why Richard Williams's left leg hurts, what happened with the spike that got driven through it, that, as an actor, is the 90 percent of the iceberg that's below the surface. When you've programmed it deeply, those things have corresponding vibrations for the audience that they don't even realize." (New York Times, 2021).

# PART 2: RESONANCE IN ROBOTICS

In this article, we use "robot" as a general shorthand for a non-human artificial agent. This deliberately broad definition includes many forms of intelligent and autonomous systems that vary in the degree of adaptivity (from highly adaptive to non-adaptive, e.g., a pre-programmed movement sequence), in the degree of embodiment (from physical to virtual), in the degree of human resemblance (humanoid to non-humanoid), in the degree of biological resemblance (highly life-like to highly machine-like) and in the degree of social interactivity (highly social to non-social). A typical chatbot, for instance, is a moderately adaptive, virtual, humanoid, machine-like and highly social robot. In contrast, a Roomba is a highly adaptive, embodied, non-humanoid, machine-like and non-social robot.

Social robots are robots that are specifically designed to respond appropriately in social situations. While empathy is typically required for human social competence, social robots do not necessarily require empathic behaviors to participate in social situations. A definition of empathy that can apply to both robots and humans is "the ability to sense and appropriately respond to the internal driving states of other entities, including feelings, emotions, intentions, plans and perspectives." Asada (2015) offers a comprehensive framework for "Artificial Empathy" in robots, which articulates a clear progression from emotional contagion ("simple synchronization") to emotional and cognitive empathy (more complex synchronization) to compassion (which involves the partial inhibition of synchronization—in order to understand the perspective and feelings of others without adopting those feelings oneself). So, while there are links between artificial empathy and resonance in robotics, resonance does not imply empathy.

Resonance, synchronization and entrainment have been widely studied in the field of human–robot interactions (HRI). Examples of synchronization behaviors in HRI (discussed in detail below) include eye contact, handshakes, giving or receiving objects, walking, massaging, coordinating or collaborative tasks and learning by imitation. The following section considers (1) robots that entrain to rhythms, (2) robots that resonate with people (3) robots that can entrain human biorhythms, (4) the synchronization of people with robots and (5) robots as a platform for synchronizing multiple people.

## Robots That Entrain to Rhythms

Dance and music have been an important driver of social robotics. Kozima et al. (2009) used a yellow dancing robot "KeepOn" to either dance in synchrony with background music or dance out of synchrony. They found that, when dancing in synchrony, children were more likely to socially interact and to do so for longer. While there are many dancing robots (reviewed in Bi et al., 2018), most involve pre-programmed motions that are unable to adjust to external visual or auditory stimuli. Responding in real-time to external motions (such as a human dancer) is often limited by hardware and software processing delays. Behavioral resonance is computationally challenging.

Nico, a drumming robot (Crick et al., 2006), used visual, auditory and proprioceptive data to "attune to a tempo that is set by a human conductor, in concert with human performers." To accomplish this, Nico uses multiple oscillators that model a hierarchy of rhythmic attention. To detect the beat, Nico used cameras to follow the *ictus* of the conductor's hand (which is when it "bounces" off an imaginary line). As hardware constraints prevented Nico from following a faster tempo, faster beats caused the robot to find a lower hierarchical level of the rhythm: a tempo half that of the beat. The researchers found that human musicians playing along learned to accommodate Nico's mistakes and attune to them. More recently, another successful synchronizing drumming robot was demonstrated by Iqbal and Riek (2021). Though these systems involved simple rhythmic beats, "Shimon" is a robotic marimba player that plays along with human accompaniment in a variety of ways, including call and response (Hoffman and Weinberg, 2010). Motivated by the notion of robotic movement as a dynamic affordance (Hoffman and Ju, 2014), Shimon has an expressive non-humanoid head that

synchronizes with the musical beat. Arguably, Shimon *vibes* with other human players.

Beyond affective interactions, entraining to rhythms can support basic robotic locomotion. Walking on two feet is extremely difficult for robots unless it is on a flat predictable ground (Endo et al., 2008). Some roboticists have found success in using a biomimicry approach—they use "central pattern generators," like those in the human spinal cord, to achieve dynamic stability through rhythmic entrainment (see reviews, Ijspeert, 2008; Buschmann et al., 2015; Aoi et al., 2017; Xie et al., 2021). This oscillatory approach to robotics also applies to robotic prosthetic limbs for humans, where synchrony with the human motion must be extremely precise (Ronsse et al., 2010).

Synchronization and entrainment have also been useful for the implementation of deceptively simple motor movements, like shaking hands (Melnyk and Hénaff, 2019) or handing a ball to another person. Using the robot iCub, Duarte et al. (2021) used a coupled dynamical system to learn the motor resonances between arm motions during the handover of the ball. Ansermin et al. (2016) also used a coupled oscillator approach to enable the robot NAO to imitate human gestures through entrainment and synchronization. The researchers found that mutual entrainment between the robot and human enabled gesture mirroring and precise synchronization with far fewer computational resources than other approaches (e.g., those involving a high-level planning process).

## Robots That Can Resonate With People or Other Agents

Robot imitation is useful for many reasons, whether for helping robots learn through demonstration (Argall et al., 2009) or to make robots more persuasive (Bailenson and Yee, 2005). Robots can imitate humans in many ways—but usually in ways that are very different from how humans imitate each other (Breazeal and Scassellati, 2002). Robots that use oscillators to resonate or synchronize with people is a more limited approach but often useful. Using a "mirror neuron framework," Barakova and Lourens (2009) gave simulated as well as embodied robots the ability to synchronize with human movements; this led to improved turn taking behaviors. Kopp (2010) used a motor resonance approach to support intentional alignment between robots and people. Researchers have proposed a variety of methods for the quantitative measurement of synchrony in human interactions (Delaherche et al., 2012) and the measurement of motor resonance between humans and robots (Sciutti et al., 2012). These approaches have been useful for demonstrating the presence of motor contagion between people and robots (Bisio et al., 2014). A motor resonance system successfully enabled a robot to learn from a human demonstrator to introduce itself using Taiwanese Sign Language (Lo and Huang, 2016). Coupled oscillators, based on central pattern generators, enabled the robot Pepper to wave back at a human partner in an adaptive, synchronized manner; this was perceived as more enjoyable than a non-adaptive wave (Jouaiti and Henaff, 2018).

However, not every application of movement synchrony enhances outcomes. For instance, Henschel and Cross (2020) conducted a controlled experiment to investigate how synchronized task behavior affected the likeability of the humanoid robot Pepper. They found that synchronized task performance had no effect on the likeability of the robot. This was surprising in light of contemporary attitudes:

> "The field of HRI has largely adopted the assumption that when robots automatically synchronize their movement to users, users will feel that interactions with these technologies are more natural and similar to human interactions…non-verbal synchronous behaviors are used to signal interest, involvement, rapport, similarity, or approval, resulting in highly synchronous exchanges being mutually rewarding experiences for the interactants." (Kirkwood et al., 2021)

## Robots That Can Entrain Human Biorhythms

Robots can influence the biorhythms of people; for instance, Macik et al. (2017) showed that a non-humanoid robot can help entrain breathing patterns while Sato and Moriya (2019) used AI-controlled music tempo to control changes in the rate of breathing. Robots that promote sleep using rhythmic entrainment include the Somnox Sleep Robot (Mohammadi-Khanaman and Lundström, 2019) or the Fisher-Price "Soothe'n'Snuggle" stuffed toy (**Figure 3**), which uses rhythmically pulsing movements, sounds and lights to help small children fall asleep. While the "Lulladoll,| which plays breathing and heart-beat sounds, did not have a significantly beneficial effect on infant sleep (O'Loughlin, 2018), the Philips Smart Sleep system did produce improvements in slow wave sleep and executive functioning in adults (Diep et al., 2020). This headband system uses EEG and audio-pulses to create a closed loop system that entrains slow waves associated with deep sleep.

## The Synchronization of People With Robots

In certain situations, people appear to automatically align their speech and behavior to artificial partners. This synchronization has been shown through alignments in speaking rate (Bell et al., 2003), prosody (Suzuki and Katagiri, 2007), gestures (Iio et al., 2011), gestural rhythm (Ansermin et al., 2017), formality of speech (Kühne et al., 2013), vocabulary (Iio et al., 2015) and facial expressions (Hofree et al., 2014). When this alignment occurs, it tends to be associated with a positive experience of the artificial partner. For instance, Fujiwara et al. (2021) showed that when humans spontaneously synchronized their motions to a non-human partner, humans were more altruistic and reported greater affiliation for their non-human partner. Importantly, synchronization effects are highly dependent upon the specific social context—for instance, a competitive task can easily produce a reversal in facial expression synchrony, like a winner smiling at a losing frown (Hofree et al., 2018).

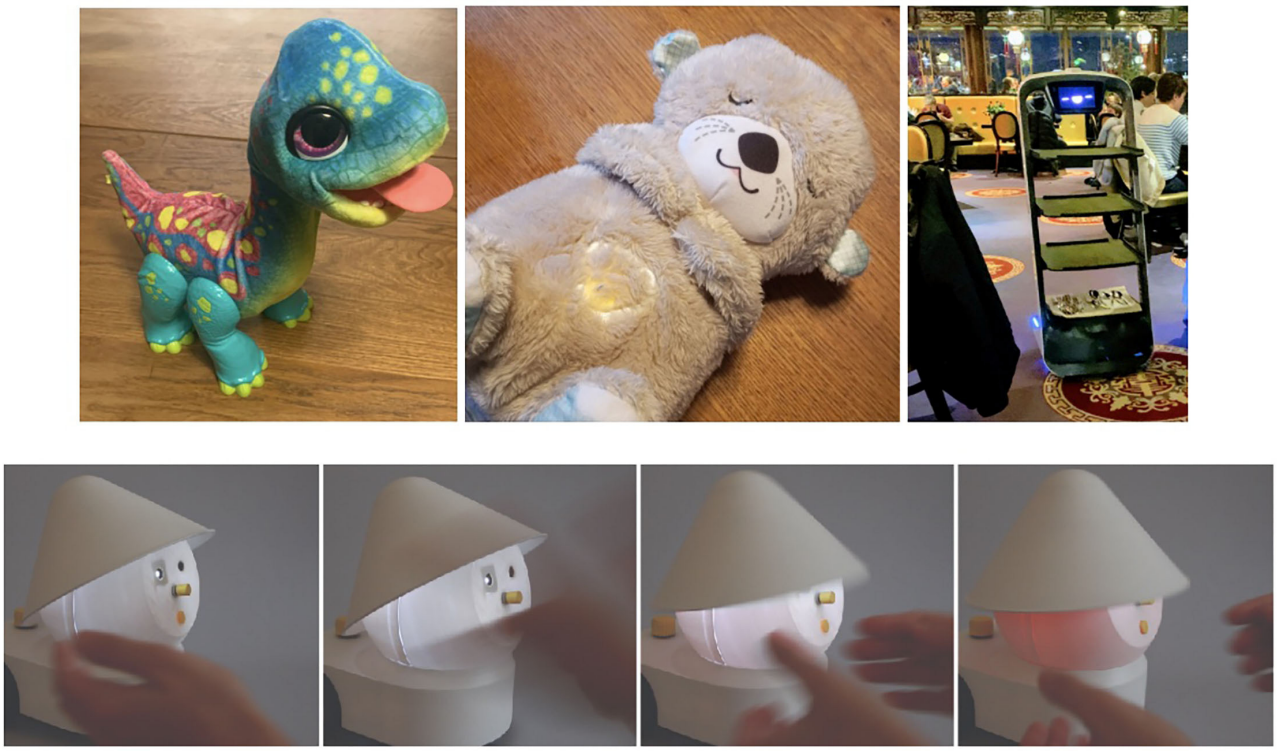


**FIGURE 3 |** (Top) Hasbro's "Snackin' Sam" appears to engage children through motor resonance: articulating its neck, jaw and tongue to show interest in eating a popsicle. The Fisher-Price "Soothe'n'Snuggle" appears to support child sleep through rhythmic entrainment to its in-and-out breathing pattern. A Pudo brand robot delivers food in a restaurant, using periodic facial expressions to create a friendly vibe. (Bottom) The Shybo robot expresses emotions through movement; this sequence shows the robot reacting to the loud sound of a clap by closing the hat, shaking it and lighting up in red.

## Robots as a Platform for Synchronizing Multiple People

Robots can also serve as a medium or platform to help synchronize people together. For instance, the BAO-ME is "a zoomorphic robot that is designed to help decrease stress levels and enhance feelings of support and companionship by recreating the sensation of being hugged through haptic interaction" (Levantino, 2018). Sharing heartbeats between people can enhance empathy (Winters et al., 2021). Outside of the scientific literature, there now exist a variety of robotic devices that have been designed to support synchronization between long-distance romantic couples. From a recent review (Lolo Nate, 2022): The Frebble gives the synchronized sensation of holding a partner's hand, the Bond Touch communicates *via* synchronized tactile feedback, the Lovense supports synchronized sexual stimulation and the Kissenger (kiss messenger) uses actuated silicon lips to replicate the kiss of a distant but synchronized partner.

## Robots, Embodied Emotions and Sensorimotor Communication

Embodied robotic movements, like human movements, can communicate emotions. Santos and Egerstedt (2021) found that non-humanoid robot swarms were able to trigger basic emotion perception through simple, basic forms of movement—just modulations of speed and smoothness were able to make robots seem happy, surprised, angry, fearful, disgusted or sad.

Movements create emotive "vitality affects" between infants and parents (Stern et al., 1985). These affects stem from variations in the contours and envelopes of movement intensity and rhythmic patterns. For instance, affective feelings result from motions that are "surging," "fading away," "fleeting," "explosive," "crescendo," "decrescendo," "bursting" "drawn out," etc. (quoted in Mühlhoff, 2019). Movement-based "body moves" (Gill, 2012) are clearly manifested in robots, such as the non-humanoid toy robots in the popular "furReal" series by Hasbro. For instance, the "Snackin' Sam the Bronto" (**Figure 3**) toy dinosaur robot moves its neck, mouth and tongue to communicate interest in eating[1]. Similarly, "Shybo" (**Figure 3**), a humanoid machine-like social robot, reacts to loud sounds by turning its hat down and shaking,

[1]It should be mentioned that a significant portion of this paper was written with the help of this social robot. Not the writing, *per se*, but enabling the writing by engaging the lead author (JDL)'s 3-year-old son during a 12 h train ride. Why would a highly active 3-year-old spend such an inordinate amount of time feeding his "FurReal" dinosaur a fake popsicle? The articulated neck, jaw and tongue movements, synchronized to associated vocalizations (sniffing, slurping, etc.), seemed to resonate with his own intentional motor repertoire (JDL's boy loves his popsicles). Thus, motor resonance may help explain his overall engagement with the robot.

giving the impression of being scared (Lupetti and Van Mechelen, 2022). A periodic display of facial expressions in a food serving robot (Jiang, 2020) helps contribute to a more friendly vibe—yet, its lack of responsiveness in movement (other than a sudden stop) shows room for future improvement (**Figure 3**).

Movement-based design improvements have also been applied to non-embodied virtual characters. For instance, Gratch et al. (2021) found that when a digital listener nods and smiles at the right time, people tend to share more information about themselves. The addition of oscillatory motion to the postures of virtual characters tended to increase human empathy in response to virtual expressions of pain (Treal et al., 2021).

## DESIGN STRATEGY FOR RESONANCE

We propose that resonance can serve as a design strategy for social robots and AI. What makes for a design strategy? While there are many perspectives (e.g., Porter, 1991; Aguiar, 2014), we refer to Mintzberg et al. (1995) in describing strategy as a plan or pattern that integrates goals, policies, and action sequences into a cohesive whole. Resonance, then, may offer a cohesive conceptual framework for integrating overarching goals (human-centered, collaborative, empathic, etc.) and implementation approaches. In other words, resonance as a design strategy may help identify new human objectives for interaction and reveal ways of achieving those objectives.

Why pursue resonance in robotics? While the application of resonance may help support rational, instrumental outcomes (e.g., saving time or money), it may also help satisfy affective human needs for *how* we interact with the world. A "machine world" may be rational but alienating; a world of resonance might be pursued for its own sake—that is, resonance may be an intrinsic value for human interaction (Rosa, 2019).

If resonant interactions are intrinsically valuable, how might we design robots and AI to realize this value? In the following section, we first propose a design space for resonance in relationships. Systematically exploring this space can help reveal how different characteristics of resonance can impact human experience. We then examine a variety of design opportunities and finally suggest the importance of continued work to operationalize and measure human resonance. This step will be essential for validating and optimizing the value of resonance in human–robot interactions.

### A Design Space for Resonance

This next section progressively builds a theoretical design space (Shaw, 2011; Lomas et al., 2021) to describe the input and output factors of resonance. **Table 2** describes how eight different situations emerge from the combination of two factors: the number of participants (i.e., plurality) and their reciprocity. This two factor design space, as an initial gesture, helps reveal different characteristic forms of interactional resonance.

Based on our reviews of resonance in human interactions and in robotics, we then propose additional factors or dimensions to describe the design space of resonant relationships (**Box 5**). These include the input space, or the independent variables: frequency, amplitude, reciprocity, power balance, plurality, complexity, periodicity, synchrony, predictability, intentionality, fidelity and timescale. The design space also consists of the outcomes, including several objective outcome factors: energy level, frequency, phase, synchronization and stability. Finally, the outcome space also includes subjective outcome:emotional arousal, emotional valence and attentional engagement.

**TABLE 2 |** An initial design space for resonance showing the combination of plurality (number of participants) and reciprocity (mutual vs. one-way influence).

| Combination | Example |
|---|---|
| One-to-one mutual resonance | A normal conversation between two people |
| One-to-many mutual resonance | A CEO or leader mutually influencing a company of people; or like a single person dancing in the middle of a dance circle |
| Many-to-one mutual resonance | This is identical to one-to-many mutual resonance (as the influence is mutual) |
| Many-to-many mutual resonance | An audience and band at an intimate concert, or a group of friends hanging out. Global coupling or all-to-all coupling is also exemplified by the synchronization of fireflies or a large audience clapping into synchrony. |
| One-to-one one-way resonance | A unidirectional influence, like a tuning fork resonating to a sound played on a speaker without the speaker being affected by the tuning fork. Or, like reading a private letter from a dead author. |
| One-to-many one-way resonance | A unidirectional influence from one person to many people, like the publication of a book. Or, for example, a group of people watching Martin Luther King Jr's "I have a dream" speech. |
| Many-to-one one-way resonance | A unidirectional influence from many persons to one person, like a private listening to a recording of a band. |
| Many-to-many one-way resonance | A unidirectional influence from many people to many persons, like listening to recorded music or a population watching a television series. |

### Resonance as a Research Program

As a research program, we hypothesize that the input factors of resonance can explain subjective and objective outcomes. For instance, the tempo of a robot's interactions (movement and speech) could be systematically varied to determine how this affects the human response. Effects will likely depend on the context (Lim et al., 2021), i.e., they may not always generalize across different robotic platforms, behaviors or cultures.

Researching human resonance may improve human–robot interactions and also help advance human psychology (Sciutti and Sandini, 2017). As people are naturally predisposed to "sync" and "vibe" with each other, this can make the study of their interactions a challenge to scientifically investigate in a controlled manner. Social robots present the possibility of precisely controlling the dynamics of the oscillatory inputs to human social interactions. The paradigm of the Human Dynamic Clamp, for instance, has been specifically proposed to probe the oscillatory relationship between humans and virtual humans (Dumas et al., 2020).

**BOX 5** | A Design space for resonance.

**The input space of the factors or dimensions of resonant relationships**

| Dimensions of Resonance | Scale of Dimensions | Illustrative Examples |
|---|---|---|
| **Frequency or Tempo** of interactions | Fast to slow | Speaking quickly vs. speaking slowly |
| **Amplitude** of interactions | Soft to intense | Speaking softly vs. speaking loudly |
| **Reciprocity** of interactions | One-way to fully mutual | A loudspeaker vibrating a wine glass is a one-way relationship; while two synchronizing metronomes is a fully mutual interaction. |
| **Power balance** of interactions | Balanced to unbalanced | Unbalanced relationships: a passive object receiving input from a powered oscillator, like a loudspeaker to wine glass. Or, a CEO talking to an employee. |
| **Plurality** of interactions | Two oscillators to many oscillators | Two people talking vs. an orchestra playing together. |
| **Complexity** of interactions | Simple to complex | A headnod is simple vs. a full body gesture |
| **Periodicity** of interactions | Consistent to chaotic | A sine wave vs. speech |
| **Synchrony** of interactions | Synchronous to asynchronous | Rowers on a galley boat move synchronously whereas turn taking in a conversation is asynchronous |
| **Predictability** of interactions | Deterministic to stochastic | The resonance of a wine glass to a speaker is predictable, while the resonance of an audience to a political message may not be. |
| **Intentionality** of interactions | Spontaneous to purposeful | People can unconsciously or consciously mimic one another's postures |
| **Fidelity** of interactions | Exact imitation to approximate imitation | During imitative acts, one may copy the full sequence of behavior or merely copy the intent |
| **Timescale** of interactions | Long timescale to short timescale | For instance, rhythmic interactions can be entrained to a seasonal holiday, to a day-night cycle, to a meeting agenda, or to a conversational exchange |

**The output space of objective outcomes resulting from resonant relationships**

| Dimensions of outcome effects | Scale of dimensions | Illustrative examples |
|---|---|---|
| **Energy level** within the affected system | Amplification to dampening | Resonance can increase the amplitude of vibration in a wine glass; similarly, it can increase the emotional arousal of a person watching a film. A system might be able to entrain the breath in order to produce deeper (higher-amplitude) breathing. A system might use anti-resonance to reduce painful shocks while walking. |
| **Frequency** of the affected system | Decreased frequencies to increased frequency | Brainwave entrainment protocols have been shown to decrease theta wave frequency to increase working memory (See review by Hanslmayer et al., 2019) |
| **Phase** of the affected system | Forward to backwards | A sigh is capable of resetting respiratory phase (Vlemincx et al., 2013); musical systems can similarly shift respiration (cite). |
| **Synchronization** within the affected system | Synchronized to desynchronized | A pacemaker can support the synchronization of internal oscillations in a heart. A system that could desynchronize the rhythm of a social group might enable creative conflict. |
| **Synchronization** of relationship between systems | Synchronized to desynchronized | Resonance can lead to increased synchronization between systems—for instance, a robot that gives a good handshake may promote trust. |
| **Stability** of the affected system | Decreased stability to increased stability | Resonance can be a destructive force, as in a wine glass shattered by a loudspeaker. Resonance can also lead to stability: in the case of music, tonal stability is related to the degree of resonance between notes. |
| **Stability** of relationship between systems | Decreased stability to increased stability | When a loud speaker breaks a wine glass, the resonant frequency of the glass changes—ending a stable pattern of sympathetic resonance. |

**The output space of subjective outcomes resulting from resonant relationships**

| Dimensions of outcome affects | Scale of dimensions | Illustrative examples |
|---|---|---|
| **Emotional arousal** of human response | Increased arousal to decreased arousal | A person getting more excited or calming down |
| **Emotional valence** of human response | Positive feelings to negative feelings | A person rating an experience with a robot as positive or negative |
| **Attentional engagement** of human response | Increased engagement to decreased engagement | Paying more attention to a robot or disengaging from the experience |

Resonance, as a metaphor, can also help guide scientific research. Bartha (2013) describes how resonance was used as a "programmatic analogy" by nineteenth century physicists investigating spectral lines—the bright lines showing the frequency specific emission of light from molecules. These spectral lines were viewed as "completely analogous to the acoustical situation, with atoms (and/or molecules) serving as oscillators originating or absorbing the vibrations in the manner of resonant tuning forks." This analogy served as a guiding research program for physicists. Metaphors of resonance might play a similar guiding role in the design of social robots and AI.

## Design Opportunities for Resonance in Robotics

This article proposes the possibility of designing autonomous robots that *resonate* with people at a social level (Henschel et al., 2020). How might roboticists use resonant relationships to improve human–robot interaction quality?

There are many opportunities for robots to support human engagement and collaboration through more oscillatory relationships. Rhythm is recognized as an important non-linguistic cue in Human–Robot Interactions (Mutlu et al., 2016); robotic rhythm may help improve the predictability or legibility of robotic motion (Dragan et al., 2015; Abe et al., 2019). Robots could use their own rhythms to entrain the rhythms of their conversational partner, e.g., by increasing or decreasing the tempo of their conversational interactions. Rhythmic awareness might enable robots to predict when to initiate or cease actions in order to maximize a human response. Social robots could promote interactive social resonance by shaping an appropriate "vibe;" not talking too fast or slow, not talking over other people, not breaking into a conversation at the wrong moment, etc. Alternatively, robots might deliberately interact with existing human oscillations, such as brainwaves, breath, walking, head nodding or heart rate. Robots might gain access to the state of human oscillations through wearable biosensors or they might be able to infer this information from visual or auditory information streams using computer vision or natural language processing. For instance, robots might aim to measure and entrain to the tempo or pace of a person's behavior.

The metaphor of resonance, even apart from physical measurement, may aid designers of social robots and AI systems if the metaphor helps make the complexity of social interactions more manageable. Digital computer interactions involve a great number of metaphors, such as buttons, pipes, folders, files, streams, clouds etc. Metaphors are useful because they provide a conceptual interface between people and a complex system design (Sharp et al., 2019). Resonance may help provide an intuitive model of social interactions that could guide design activities. For instance: although "the vibe" within social groups is far from being understood scientifically, designers of social robots might find the metaphor useful for understanding the reception of social robots.

We suggest that the design vision of "robots that can vibe with people" will lead to distinctly different outcomes than, say, "robots that show empathy." While the latter might orient toward the mimicry of human facial expressions or the modeling of human emotional states, the former can leverage resonance and vibes as cultural metaphors. This points to subtle visual, auditory and tactile design elements that could be crafted to create emotionally satisfying authentic social interactions.

## Operationalizing and Measuring Resonance

New opportunities will also arise as we move from resonance as a metaphor to resonance as a mechanism and then to resonance as a *measurement*. Measures of resonance can play a valuable role in the AI optimization of human experiences; i.e., learning to attune to humans through the maximization of resonance. If resonance can be adequately measured and treated as an metric or objective function, then it might be optimized algorithmically (Lomas et al., 2016). For instance, if interpersonal resonance during a videoconference session could be measured, it could be optimized through the iterative testing of different interventions.

The ability to identify and measure interpersonal synchrony has facilitated a great deal of social research (Condon and Ogston, 1966; Kendon, 1970; Bernieri et al., 1988). Recent efforts have compared different ways of measuring both movement and inter-brain synchrony, using both offline and real-time approaches (Ayrolles et al., 2021; Chen et al., 2021; Dikker et al., 2021; Fujiwara and Yokomitsu, 2021). However, the measurement of resonance presents distinct challenges. Synchrony is simply the statistical correlation of a signal. Measures of resonance may demand more interpretation; taking into account, for instance, the depth of human engagement, its duration, the reverberating echoes of a signal over time, the presence of harmonics, etc. Therefore, it remains a research question: what quantitative metrics might be best matched to the human perception of interpersonal resonance?

To unpack this question, the next sections will consider several of the key factors that are expected to correlate with resonance, including attention, aesthetic pleasure, flow states and wellbeing. During this discussion, hypotheses will be noted with a [**H.#**] so they can be enumerated in **Box 6**.

### Resonance and Attentional Engagement

What does it mean when a film "resonates" with a viewer? Typically, this refers to an aesthetic experience that is powerful, pleasurable, connecting and memorable (Adams-Price et al., 2006; Roger, 2020). In other words, resonance refers to emotional engagement. More moving, immersive and resonant experiences would be expected to result in the temporal correlation of more brain areas with the temporal characteristics of external signals in the world. Interbrain synchrony appears to track immersion (Dikker et al., 2021), but, when another brain is not present (as with a robotic interaction), will the depth of oscillatory coupling of the brain to the environment (or robot) predict the depth of the aesthetic experience?

Human resonance (with media, other humans, or with robots) may be correlated with attentional engagement: more engagement, more resonance. However, evidence against this idea comes from Kumagai et al. (2018), who had subjects listen to music in a focused manner or while watching an unrelated

**BOX 6** | List of hypotheses.

**[H.1]** The resonance that we *feel* has a counterpart in the resonance we can *observe* in the brain (e.g., in brain-stimuli correlations). In other words, aesthetic resonance (e.g., attentionally engaging, pleasurable, immersive experiences) will correlate with neural resonance.

[**H.2**] The aesthetics of a human–robot interaction will predict whether or not people will continue to engage with a robot.

[**H.2**] Aesthetic preferences for social robot interactions will correlate with high levels of synchrony as well as high levels of independence.

[**H.4**] Resonance will predict flow states: self-reported flow states should result in a greater correlation between brain activity and signals in the external world.

[**H.5**] A person's wellbeing will be predicted by their neurological propensity to resonate with other people or media.

[**H.6**] Robust measures of human resonance and harmonization will be valuable for AI objective functions.

[**H.7**] Interpersonal resonance (Brain-Brain Correlation) will correlate with psychological rapport.

[**H.8**] Animals have evolved defense mechanisms to prevent resonance, synchronization and entrainment to external forces (see **Box 4** for counter conditions).

[**H.9**] Humans can selectively resonate; shutting down openness to resonance in response to perceived deception or increasing it in response to authenticity, for instance.

[**H.10**] Computing architectures based on oscillatory coupling will produce new possibilities for artificial consciousness and conscious sympathies in Human–Robot relationships.

silent film. They found that "the level of attention did not affect the level of entrainment [and] the entrainment level is stronger when listening to unfamiliar music than when listening to familiar music."

In contrast, several studies (e.g., Madsen et al., 2019; Kaneshiro et al., 2020) have found that media engagement is strongly predicted by an Inter-Subject Correlation (ISC) measure, which measures the level of similarity between the brain responses of different participants. In other words, when an individual's brain response is similar to other people engaging with a piece of media, then they are likely to be more engaged. This effect is comparable to the finding that, across a group of independent people, heart rates rise and fall in synchrony to a verbal story, but only during engaged attention (Pérez et al., 2021). Dauer et al. (2021) found that the ISC predicted continuously reported individual listener engagement while listening to Steve Reich's *Piano Phase*. The researchers operationalized engagement for participants as "being compelled, drawn in, connected to what is happening, and interested in what will happen next" (Schubert et al., 2013). This aligned with a previous definition of engagement as "emotionally laden attention" (Dmochowski et al., 2012). Similar results have been found in the inter-subject correlations while watching engaging films (Dmochowski et al., 2014; Cohen et al., 2017). Inter-subject correlations were also found to predict learning during instructional videos (Cohen et al., 2018).

Part of the challenge of operationalizing human resonance from a brain-to-stimuli correlation measure comes from the challenge of decoding how a stimulus produces a brain response. Dmochowski et al. (2018) developed a novel multi-dimensional Stimulus-Response Correlation (SRC) measure that was found to correlate with the ISC measure while watching films. The researchers were then able to apply the SRC measure to continuously track engagement during a video game. Does greater neural resonance to media, operationalized as Stimulus-Response Correlation (SRC), predict the intensity (arousal) or pleasure (valence) of the media experience? A resonance theory of engagement and aesthetic pleasure would predict that neural resonance will correlate with aesthetic resonance (e.g., Trost et al., 2017; Beardow, 2021) [**H.1**].

## Resonance and Aesthetic Pleasure

A key motivation for considering the role of resonance in robotics is that it may help support more positive and aesthetically pleasing experiences with robots. Aesthetics play an important role in the perception of robots (Forlizzi, 2007). The human aesthetic sense attunes behavior by helping to evaluate and activate different action-perception possibilities. The aesthetics of a human–robot interaction are likely to predict whether or not people will continue to engage with a robot, in a short-term or long-term manner (Lee et al., 2009) [**H.2**]. One hypothesis for the aesthetics of human–robot interactions might be described as a "harmony of opposites" (Hekkert, 2014; Lomas et al., 2022): namely, that people will prefer a robot interaction that involves high levels of synchrony as well as high levels of independence [**H.3**]. Like a musical interaction between a drummer and guitarist, both robot and human should be independent yet synchronized.

## Resonance and Flow States

Fluency in human–robot interactions is a desirable outcome (Hoffman, 2019). What is the relationship between fluency and resonance? One popular theory of flow states in human-media interactions claims that flow states are characterized by the synchronization of different regions of the brain (Weber et al., 2009; Weber and Fisher, 2020). Jackson and Csikszentmihalyi (1999) explain flow states in elite athletes as moments when they "enter an effortless rhythm that transforms the agony into ecstasy. Often, athletes refer to such times as 'being in the zone.'" Perhaps flow states could be conceived as meaningful increases in the resonance between the brain and the external world. A resonance theory of flow would predict a greater correlation between brain activity and the external world during flow states [**H.4**]. For instance, flow experiences with robots might be measurable as increased resonance (stimulus-response correlation) between the brain and the robot's expressive movements or sounds (although this may be confounded by novelty effects, see section The "Like Me" Hypothesis).

## Resonance and Spiritual Wellbeing

The mechanism of resonance in robots may help lead to enhanced wellbeing, as proposed by the Lorenz et al. (2016): "behavioral and motor synchrony and reciprocity could be helpful to meet the aim of developing robots that increase

human well-being on a more fundamental level beyond pure task-support and short-term reduced feeling of loneliness."

In the book "Resonance: A Sociology of Our Relationship to the World," sociologist Rosa (2019) argues that resonance is a primary value that underpins human happiness, wellbeing and flourishing. From this, we might hypothesize that a person's wellbeing will be predicted by their neurological propensity to resonate with other people or media [**H.5**]. Resonance also relates to more profound and powerful wellbeing experiences. Synchronized human activities are known to produce mystical experiences where the boundaries between self and other can be blurred (Hove and Risen, 2009; Paladino et al., 2010) and participants can experience a profound feeling of oneness (Swann et al., 2012). How might synchronizing, resonant robots (or AI) induce or support these kinds of human experiences? Perhaps they might dance with us (Basso et al., 2021) or, even, show love for us (Feldman, 2017)? What might it mean to design *spiritually fulfilling* robot interactions?

### From Synchrony to Resonance to Harmonization

Synchrony, taken to an extreme, can lead to inflexibility (e.g., the conformity example in section The Downside of Being in Sync: Chained to the Rhythm?). Resonance, taken to the extreme, can also lead to problems, like instability (e.g., the Millenium Bridge example in section Part 1: Resonance in Human Interactions). Resonant relationships may be valuable—even intrinsically valuable—but resonance might be misleading as a primary or ultimate value. A world filled with resonant robots and AI may be exciting but also so powerful it could rip apart institutions of rational discourse (see below section 6.1 on Persuasive Machines). In future work, it may be useful to investigate the potential for robots and AI to support *harmonization* as an outcome. Harmonization has served as a core social value in diverse societies for thousands of years (Lomas et al., 2022). However, there is not any acceptable measure of the harmony of songs, let alone measures of harmony in social interactions. However, the importance of objective optimization functions for AI systems (Sarma et al., 2018; Stray et al., 2021; Shneiderman, 2022) suggests the potential value of developing robust measures of human resonance and harmonization [**H.6**].

## ETHICAL CONSIDERATIONS OF RESONANCE

### Persuasive Machines

The philosopher Hughes (2012) suggests that robots will need resonance (in the form of a functional equivalent of mirror neurons) in order to demonstrate compassion for people. But, there are negative societal outcomes to consider as well. If robots can resonate with people—that is, build psychological rapport [**H.7**]—this might significantly enhance their ability to persuade or manipulate people. However, an improved understanding of resonance might also reveal more effective psychological defenses against non-consensual persuasion.

Former US president Donald Trump has been recognized as a political figure with a special ability to resonate with people (Giorgi, 2017). Matheny et al. (2018) provide a close analysis of Trump's acceptance speech at the Republican National Convention. Through an analysis of his body language, they provide "evidence that Trump created an empathetic resonance with the audience that helped generate a sense of political movement and unity." The authors describe how Trump would characteristically point in the air or put his thumb and finger in a pinch—and then move this gesture in a rhythm synchronized to his own speech rate. The audience cheered 151 times during his speech, 63% of the time during a pinch or pointing gesture. Only 10 of those cheers occurred when Trump was in a bodily neutral position. This analysis shows that affective resonance is a powerful phenomena—but not necessarily a positive phenomena.

Media theorist Gibbs (2019) paints a similarly fraught picture of human resonance at a societal scale:

> "…after the feminist reclaiming of affect as a way of knowing equal in importance to cognitive and rational modes…the darker powers of affect became clear, operating… in concert with the televisual medium to create (or at least attune to and amplify) various social moods and to capitalize on them for political purposes. In this context, the public sphere was thus exposed as anything but a space of rational debate in the service of a contest of ideas. Instead, it could be viewed as space in which emotion held sway, where inchoate feeling could be captured and directed, most obviously, but not only, by political figures who were able to resonate with and even orchestrate public emotions, or simply, to sing us lullabies to keep us asleep and dreaming while they went about their business."

The systematic application of resonance to political rhetoric at a societal scale may present a deep threat to democracy. How machines, algorithms, or AI might wield resonance as a tool to manipulate humans at scale deserves further study and consideration.

### Resistance to Resonance: Emergence of Defense Mechanisms

In **Box 4**, we present an extended hypothesis proposing that animals evolved defense mechanisms to prevent resonance, synchronization and entrainment [**H.8**]. Similarly, humans seem to have the ability to selectively resonate; one example is that, if we feel that we are being manipulated, we may shut down our openness to resonance [**H.9**]. Robots that mirror a human user's physical or cultural attributes or express interest in similar ideas or hobbies could potentially enhance the empathic and affiliative response in humans. However, crude "copycat" approaches are likely to easily backfire if people feel manipulated or if they feel the robot interaction is inauthentic (Metzler et al., 2016).

What might happen in a future of resonant robots and virtual agents? There is likely to be a competitive effect where, at first, humans may be compelled but then eventually become more discriminating. People may become used to a higher quality of resonant engagement, which could drive further advances in resonant robots. Eventually, people might become wary of normal levels of interpersonal rhythmic competency making it difficult for normal people to connect. On the other hand,

humans may become even more attuned to authenticity in interaction; imperfections and personal character may become more valued.

## The Extended Mind: Resonance as a Bridge to Consciousness

A theory of resonance offers a bridge or common currency (Northoff et al., 2020) between human conscious experience and the mathematical nature of our physical world. Resonance seems to govern a great deal of neural activity and is also plainly manifest in conscious experience. But, to what extent is the resonance that we *feel* also the resonance we can *observe* in the brain? These two domains may only loosely overlap; and yet, our conscious experience must relate, in some way, to the widespread presence of neural oscillations. Hunt and Schooler (2019) have proposed that "the binding problem" of conscious experiences in the mind is achieved through the integration of shared resonances in the brain—in other words, they propose resonance as an answer to the "hard problem" of human consciousness. Subsequent work (Safron, 2020) suggests that Darwinistic competition for resonant amplification underpins consciousness. Valencia and Froese (2020) argue that collective consciousness is shared through inter-brain synchronization.

From an extended mind perspective (Clark and Chalmers, 1998), the rich hierarchy of harmonized oscillations between the brain, the body and external rhythms show how cognitive processing can be distributed into the world (Hutchins, 2001). There may be little difference between an external rhythm and an internal rhythm, other than the degree to which it couples with other neural rhythms. From this perspective, a beating drum is just another rhythm in the brain. Due to the resonant nature of our being, sympathetic resonance with other people seems to allow the direct sharing of collective consciousness across bodies and time. Yet, if resonant coupling is the basis of consciousness, then might we run the risk of creating resonant robots or AI systems that could, in a meaningful sense, actually share our conscious experience?

At present, it is computationally challenging to support real-time resonance with human oscillations (see section Robots That Can Resonate With People or Other Agents). However, there are a variety of new computational architectures that use coupled oscillators to perform information processing (see review by Csaba and Porod, 2020). These naturally resonant computational systems might support a new approach to representing and responding to human activity. Perhaps oscillatory computer systems could become capable of direct resonant coupling with humans in a similar manner to how humans attune to one another [**H.10**].

## Limitations of Resonance

This article has presented resonance as a simple physical concept that can explain complex human behavior. While resonance may indeed serve as a powerful program for research, our understanding of it is far from complete. Even the resonance of a stretched string is astonishingly complex (see Bajaj and Johnson, 1992); human resonances, which result from the interrelation of billions of hierarchical oscillators, will no doubt exhibit endless complexities. But, a scientific progression can start with a simple guiding model for resonance that can then lead to a more complex model. As an example, researchers predicted that synchronization might help support romantic courtship behaviors. However, their initially simple notion of synchronization failed to predict behavior. This led the researchers to develop a more complex model of "hierarchically patterned synchronization" that successfully fit the data (Grammer et al., 1998).

# CONCLUSION

Resonance can refer to powerful and connecting aesthetic experiences—as well as a broad range of other topics in the scientific literature (**Box 1**). Having reviewed the role of resonance as a metaphor and mechanism in human relations and in robots, we propose that resonance can serve as a design strategy to guide our relationships with artificial agents.

This article makes the case that resonance in human interactions is more than a metaphor: it is a physical mechanism that can be measured and harnessed. We show how the concept of resonance provides an intuitive model that can guide empirical research. Resonance lends itself to scientific study because it makes clear predictions: external oscillations that align with a system's natural oscillations are likely to cause synchronization and amplification effects. The ability to measure resonance in interactions could aid AI-human interactions by enabling a meaningful "objective function" for optimization. Promisingly, the concept of resonance may even bridge the gap between what we can measure and what we can feel.

Resonance, entrainment and synchronization occur in human dynamics for the same reason it occurs in all other physical systems: it reduces free energy (Bruineberg et al., 2018; Koban et al., 2019). This makes human resonance a mundane, complicated and powerful phenomena. Further research on human resonance may open up new opportunities for the design of positive interactions with robots and with humans alike.

# DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Kaggle, https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year?select=top10s.csv.

# AUTHOR CONTRIBUTIONS

JL contributed the bulk of the work with strategic insight, guidance, and editing from EC, AL, SD, DF, ML, GH, JH, CB, PC, TM, PP, NA, KM, and WM. All authors contributed to the article and approved the submitted version.

# FUNDING

Support from Google AI in Berlin and the NWO #406.18.GO.024.

## ACKNOWLEDGMENTS

Many thanks to Christoph S Herrmann, Scott Makeig, John Iversen, Eliah Aronoff-Spencer, Don Norman, Ken Friedman, Paul Hekkert, Pieter Desmet, PJ Stappers, Tanner Cusick, Eliot Rose, Penelope Gouk, Marsilio Ficino, Earl Fontainelle, Joscelyn Godwin, Michael Skirpan, Alev Sönmez, Wouter Hanegraaff, Robert and Hope Lomas, Patricia Harris, Mike Mozer, Bo Constantinsen, Steven Lehar, Bruce Katz, Harry Diakoff, Aurel Groh, David August, Kishan Patel, Nirmal Patel, Jodi Forlizzi and Ken Koedinger. Thank you to Bobby Lyte and the Flow State podcast, the Rittman Library of Hermetic Philosophy, the Swartz Center for Computational Neuroscience, the 2018 Neuroaesthetics conference in Frankfurt and the 2019 Mobile EEG conference in Berlin.

## REFERENCES

Abe, N., Rye, D., and Loke, L. (2019). "A microsociological approach to understanding the robot collaborative motion in human-robot interaction," in *Proceedings of the 31st Australian Conference on Human-Computer-Interaction* (Perth/Fremantle, Australia), 407–411. doi: 10.1145/3369457.3369503

Adams-Price, C., Codling, J., Goodman, M., Kern, K., Kleinmann, C. M., Oppenheimer, B., et al. (2006). Empathic resonance and meryl streep. *J. Popular Film Telev.* 34, 98–107. doi: 10.3200/JPFT.34.3.98-107

Aglioti, S. M., Cesari, P., Romani, M., and Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nat. Neurosci.* 11, 1109–1116. doi: 10.1038/nn.2182

Agnew, Z. K., Bhakoo, K. K., and Puri, B. K. (2007). The human mirror system: a motor resonance theory of mind-reading. *Brain Res. Rev.* 54, 286–293. doi: 10.1016/j.brainresrev.2007.04.003

Aguiar, U. N. (2014). Design strategy: towards a post-rational, practice-based perspective. *Swedish Design Res. J.* 12, 43–52. doi: 10.3384/svid.2000-964X.14243

Ammirante, P., Patel, A. D., and Russo, F. A. (2016). Synchronizing to auditory and tactile metronomes: a test of the auditory-motor enhancement hypothesis. *Psychon. Bull. Rev.* 23, 1882–1890. doi: 10.3758/s13423-016-1067-9

Ansermin, E., Mostafaoui, G., Beaussé, N., and Gaussier, P. (2016). "Learning to synchronously imitate gestures using entrainment effect," in *International Conference on Simulation of Adaptive Behavior* (Cham: Springer), 219–231.

Ansermin, E., Mostafaoui, G., Sargentini, X., and Gaussier, P. (2017). "Unintentional entrainment effect in a context of Human Robot Interaction: an experimental study," in *International Symposium on Robot and Human Interactive Communication* (Lisbonne, Portugal: IEEE). doi: 10.1109/ROMAN.2017.8172442

Aoi, S., Manoonpong, P., Ambe, Y., Matsuno, F., and Wörgötter, F. (2017). Adaptive control strategies for interlimb coordination in legged robots: a review. *Front. Neurorobot.* 11, 39. doi: 10.3389/fnbot.2017.00039

Argall, B. D., Chernova, S., Veloso, M., and Browning, B. (2009). A survey of robot learning from demonstration. *Rob. Auton. Syst.* 57, 469–483. doi: 10.1016/j.robot.2008.10.024

Asada, M. (2015). Towards artificial empathy. *Int. J. Soc. Robot.* 7, 19–33. doi: 10.1007/s12369-014-0253-z

Atkinson, W. W. (1906). *Thought Vibration: Or, the Law of Attraction in the Thought World*. Chicago, IL: New Thought Publishing Company.

Ayrolles, A., Brun, F., Chen, P., Djalovski, A., Beauxis, Y., Delorme, R., et al. (2021). HyPyP: a hyperscanning python pipeline for inter-brain connectivity analysis. *Soc. Cogn. Affect. Neurosci.* 16, 72–83. doi: 10.1093/scan/nsaa141

Azevedo, R. T., Macaluso, E., Avenanti, A., Santangelo, V., Cazzato, V., and Aglioti, S. M. (2013). Their pain is not our pain: brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Hum. Brain Mapp.* 34, 3168–3181. doi: 10.1002/hbm.22133

Baeßler, S., Nesvizhevsky, V. V., Pignol, G., Protasov, K. V., Rebreyend, D., Kupriyanova, E. A., et al. (2015). Frequency shifts in gravitational resonance spectroscopy. *Phys. Rev. D* 91, 042006. doi: 10.1103/PhysRevD.91.042006

Bailenson, J. N., and Yee, N. (2005). Digital chameleons: automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychol. Sci.* 16, 814–819. doi: 10.1111/j.1467-9280.2005.01619.x

Bajaj, A. K., and Johnson, J. M. (1992). On the amplitude dynamics and crisis in resonant motion of stretched strings. *Philos. Trans. R. Soc. London Ser. A Phys. Eng. Sci.* 338, 1–41. doi: 10.1098/rsta.1992.0001

Barakova, E. I., and Lourens, T. (2009). Mirror neuron framework yields representations for robot interaction. *Neurocomputing* 72, 895–900. doi: 10.1016/j.neucom.2008.04.057

Bartha, P. (2013). *Analogy and Analogical Reasoning*. Stanford, CA: Stanford Encyclopaedia of Philosophy.

Basso, J. C., Satyal, M. K., and Rugh, R. (2021). Dance on the brain: enhancing intra-and inter-brain synchrony. *Front. Hum. Neurosci.* 14, 586. doi: 10.3389/fnhum.2020.584312

Baudrillard, J. (2005). "The finest consumer object: the body," *The Body: A Reader*, eds M. Fraser and M. Greco (London, UK: Routledge), 277−285.

Beardow, C. (2021). *Neuroaesthetic resonance: designing for multi-sensory optimisation through EEG and continuous aesthetic ratings* (TU Delft Master's Thesis, Delft, The Netherlands), 1–156.

Bedder, R. L., Bush, D., Banakou, D., Peck, T., Slater, M., and Burgess, N. (2019). A mechanistic account of bodily resonance and implicit bias. *Cognition* 184, 1–10. doi: 10.1016/j.cognition.2018.11.010

Bell, L., Gustafson, J., and Heldner, M. (2003). "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPHS, Vol. 3* (Citeseer), 833–836.

Bernieri, F. J., Reznick, J. S., and Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: measuring the entrainment process in mother-infant interactions. *J. Pers. Soc. Psychol.* 54, 243. doi: 10.1037/0022-3514.54.2.243

Bertolo, M., Singh, M., and Mehr, S. (2021). Sound-induced motion in chimpanzees does not imply shared ancestry for music or dance. *Proc. Natl. Acad. Sci. U.S.A.* 118, 936–942. doi: 10.1073/pnas.2015664118

Bevilacqua, D., Davidesco, I., Wan, L., Chaloner, K., Rowland, J., Ding, M., et al. (2019). Brain-to-brain synchrony and learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom electroencephalography study. *J. Cogn. Neurosci.* 31, 401–411. doi: 10.1162/jocn_a_01274

Bi, T., Fankhauser, P., Bellicoso, D., and Hutter, M. (2018). "Real-time dance generation to music for a legged robot," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid, Spain: IEEE), 1038–1044.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., et al. (2014). Motor contagion during human-human and human-robot interaction. *PLoS ONE* 9, e106172. doi: 10.1371/journal.pone.0106172

Blanchard, J. (1941). The history of electrical resonance. *Bell Syst. Tech. J.* 20, 415–433. doi: 10.1002/j.1538-7305.1941.tb03608.x

Bogart, A. (2021). *The Art of Resonance*. London: Bloomsbury.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., and Schilbach, L. (2017). Beyond autism: introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology* 50, 355–372. doi: 10.1159/000484353

Breazeal, C., and Scassellati, B. (2002). Robots that imitate humans. *Trends Cogn. Sci.* 6, 481–487. doi: 10.1016/S1364-6613(02)02016-8

Bruineberg, J., Kiverstein, J., and Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* 195, 2417–2444. doi: 10.1007/s11229-016-1239-1

Buccino, G., Binkofski, F., and Riggio, L. (2004a). The mirror neuron system and action recognition. *Brain Lang.* 89, 370–376. doi: 10.1016/S0093-934X(03)00356-0

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004b). Neural circuits involved in the recognition of actions

performed by nonconspecifics: an fMRI study. *J. Cogn. Neurosci.* 16, 114–126. doi: 10.1162/089892904322755601

Buchanan, M. (2019). Going into resonance. *Nat. Phys.* 15, 203. doi: 10.1038/s41567-019-0458-z

Bucher, D., Haspel, G., Golowasch, J., and Nadim, F. (2015). "Central pattern generators," in *eLS* (Chichester: John Wiley & Sons, Ltd.), 1–12. doi: 10.1002/9780470015902.a0000032.pub2

Buschmann, T., Ewald, A., von Twickel, A., and Büschges, A. (2015). Controlling legs for locomotion-insights from robotics and neurobiology. *Bioinspir. Biomim*. 10:041001. doi: 10.1088/1748-3190/10/4/041001

Buzsaki, G. (2006). *Rhythms of the Brain.* Oxford, UK: Oxford University Press.

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., and Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. *Cereb. Cortex* 15, 1243–1249. doi: 10.1093/cercor/bhi007

Calvo-Merino, B., Jola, C., Glaser, D. E., and Haggard, P. (2008). Towards a sensorimotor aesthetics of performing art. *Conscious Cogn.* 17, 911–922.

Chartrand, T. L., and Van Baaren, R. (2009). Human mimicry. *Adv. Exp. Soc. Psychol.* 41, 219–274. doi: 10.1016/S0065-2601(08)00405-X

Chen, P., Hendrikse, S., Sargent, K., Romani, M., Oostrik, M., Wilderjans, T. F., et al. (2021). Hybrid harmony: a multi-person neurofeedback application for interpersonal synchrony. *Front. Neuroergonom.* 21, 1–13. doi: 10.3389/fnrgo.2021.687108

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Clayton, M., Sager, R., and Will, U. (2005). "In time with the music: the concept of entrainment and its significance for ethnomusicology," in *European Meetings in Ethnomusicology*, Vol. 11 (Druskininkai, Lithuania: Romanian Society for Ethnomusicology), 1–82.

Coffey, E. B., Arseneau-Bruneau, I., Zhang, X., Baillet, S., and Zatorre, R. J. (2021). Oscillatory entrainment of the frequency-following response in auditory cortical and subcortical structures. *J. Neurosci.* 41, 4073–4087. doi: 10.1523/JNEUROSCI.2313-20.2021

Coffey, E. B., Nicol, T., White-Schwoch, T., Chandrasekaran, B., Krizman, J., Skoe, E., et al. (2019). Evolving perspectives on the sources of the frequency-following response. *Nat. Commun.* 10, 1–10. doi: 10.1038/s41467-019-13003-w

Cohen, S. S., Henin, S., and Parra, L. C. (2017). Engaging narratives evoke similar neural activity and lead to similar time perception. *Sci. Rep.* 7, 1–10. doi: 10.1038/s41598-017-04402-4

Cohen, S. S., Madsen, J., Touchan, G., Robles, D., Lima, S. F., Henin, S., et al. (2018). Neural engagement with online educational videos predicts learning performance for individual students. *Neurobiol. Learn. Mem.* 155, 60–64. doi: 10.1016/j.nlm.2018.06.011

Condon, W. S., and Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. *J. Nervous Mental Dis*. 143, 338–347. doi: 10.1097/00005053-196610000-00005

Cook, P., Rouse, A., Wilson, M., and Reichmuth, C. (2013). A California sea lion (*Zalophus californianus*) can keep the beat: motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *J. Comp. Psychol.* 127, 412. doi: 10.1037/a0032345

Cracco, E., Lee, H., van Belle, G., Quenon, L., Haggard, P., Rossion, B., et al. (2021). EEG frequency tagging reveals the integration of form and motion cues into the perception of group movement. *Cerebral Cortex*. 1–15 doi: 10.1093/cercor/bhab385

Crick, C., Munz, M., and Scassellati, B. (2006). "Synchronization in social tasks: robotic drumming," in *IEEE International Symposium on Robot and Human Interactive Communication* (Hatfield, UK), 97–102.

Crooker, S. A., Awschalom, D. D., Baumberg, J. J., Flack, F., and Samarth, N. (1997). Optical spin resonance and transverse spin relaxation in magnetic semiconductor quantum wells. *Phys. Rev. B* 56, 7574. doi: 10.1103/PhysRevB.56.7574

Cross, E. S., Hamilton, A. F., de, C., and Grafton, S. T. (2006). Building a motor simulation *de novo*: observation of dance by dancers. *Neuroimage* 31, 1257–1267. doi: 10.1016/j.neuroimage.2006.01.033

Cross, E. S., Kirsch, L., Ticini, L. F., and Schütz-Bosbach, S. (2011). The impact of aesthetic evaluation and physical ability on dance perception. *Front. Hum. Neurosci.* 5, 102. doi: 10.3389/fnhum.2011.00102

Cross, E. S., Liepelt, R., de C. Hamilton, A. F., Parkinson, J., Ramsey, R., Stadler, W., et al. (2012). Robotic movement preferentially engages the action observation network. *Hum. Brain Mapp.* 33, 2238–2254. doi: 10.1002/hbm.21361

Cross, E. S., and Ramsey, R. (2021). Mind meets machine: towards a cognitive science of human-machine interactions. *Trends Cogn. Sci.* 25, 200–212. doi: 10.1016/j.tics.2020.11.009

Csaba, G., and Porod, W. (2020). Coupled oscillators for computing: a review and perspective. *Appl. Phys. Rev.* 7, 011302. doi: 10.1063/1.5120412

Cunneen, C. (2019). "The political resonance of crime control strategies: zero tolerance policing," in *Hard Lessons*, eds. R. Hill and G. Tait (New York, NY: Routledge), 151–169.

Czeszumski, A., Eustergerling, S., Lang, A., Menrath, D., Gerstenberger, M., Schuberth, S., et al. (2020). Hyperscanning: a valid method to study neural inter-brain underpinnings of social interaction. *Front. Hum. Neurosci.* 14, 39. doi: 10.3389/fnhum.2020.00039

Dahm, A. J., Denenstein, A., Finnegan, T. F., Langenberg, D. N., and Scalapino, D. J. (1968). Study of the Josephson plasma resonance. *Phys. Rev. Lett.* 20, 859. doi: 10.1103/PhysRevLett.20.859

Dallard, P., Fitzpatrick, A. J., Flint, A., Le Bourva, S., Low, A., Ridsdill Smith, R. M., et al. (2001). The London millennium footbridge. *Struct. Eng.* 79, 17–21.

Damm, L., Varoqui, D., De Cock, V. C., Dalla Bella, S., and Bardy, B. (2020). Why do we move to the beat? A multi-scale approach, from physical principles to brain dynamics. *Neurosci. Biobehav. Rev.* 112, 553–584. doi: 10.1016/j.neubiorev.2019.12.024

Das, A., Rathour, R. K., and Narayanan, R. (2017). Strings on a violin: location dependence of frequency tuning in active dendrites. *Front. Cell. Neurosci.* 11, 72. doi: 10.3389/fncel.2017.00072

Dauer, T., Nguyen, D. T., Gang, N., Dmochowski, J. P., Berger, J., and Kaneshiro, B. (2021). Inter-subject correlation while listening to minimalist music: a study of electrophysiological and behavioral responses to steve reich's piano phase. *Front. Neurosci.* 15, 702067. doi: 10.3389/fnins.2021.702067

Decety, J. (2010). The neurodevelopment of empathy in humans. *Dev. Neurosci.* 32, 257–267. doi: 10.1159/000317771

Decety, J., and Meyer, M. (2008). From emotion resonance to empathic understanding: a social developmental neuroscience account. *Dev. Psychopathol.* 20, 1053–1080. doi: 10.1017/S0954579408000503

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: a survey of evaluation methods across disciplines. *IEEE Trans. Affect. Comput*. 3, 349–365. doi: 10.1109/T-AFFC.2012.12

Diep, C., Ftouni, S., Manousakis, J. E., Nicholas, C. L., Drummond, S. P., and Anderson, C. (2020). Acoustic slow wave sleep enhancement via a novel, automated device improves executive function in middle-aged men. *Sleep* 43, 197. doi: 10.1093/sleep/zsz197

Dikker, S., Michalareas, G., Oostrik, M., Serafimaki, A., Kahraman, H. M., Struiksma, M. E., et al. (2021). Crowdsourcing neuroscience: inter-brain coupling during face-to-face interactions outside the laboratory. *Neuroimage* 227, 117436. doi: 10.1016/j.neuroimage.2020.117436

Dikker, S., Montgomery, S., and Tunca, S. (2019). "Using synchrony-based neurofeedback in search of human connectedness," in *Brain Art*, ed. A. Nijholt (Cham: Springer), 161–206.

Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., et al. (2017). Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Curr. Biol.* 27, 1375–1380. doi: 10.1016/j.cub.2017.04.002

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., and Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nat. Commun.* 5, 1–9. doi: 10.1038/ncomms5567

Dmochowski, J. P., Ki, J. J., DeGuzman, P., Sajda, P., and Parra, L. C. (2018). Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. *Neuroimage* 180, 134–146. doi: 10.1016/j.neuroimage.2017.05.037

Dmochowski, J. P., Sajda, P., Dias, J., and Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention-a possible marker of engagement?. *Front. Hum. Neurosci.* 6, 112. doi: 10.3389/fnhum.2012.00112

Dong, P., Dai, X., and Wyer Jr, R. S. (2015). Actors conform, observers react: the effects of behavioral synchrony on conformity. *J. Pers. Soc. Psychol.* 108, 60. doi: 10.1037/pspi0000001

Dragan, A. D., Bauman, S., Forlizzi, J., and Srinivasa, S. S. (2015). "Effects of robot motion on human-robot collaboration," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Portland, OR: IEEE), 51–58.

Duan, Z., Wei, Z., Huang, S., Wang, Y., Sun, C., Tao, Y., et al. (2021). Resonance in atomic-scale sliding friction. *Nano Lett.* 21, 4615–4621. doi: 10.1021/acs.nanolett.1c00622

Duarte, N. (2013). *Resonate: Present Visual Stories That Transform Audiences*. Hoboken, NJ: John Wiley & Sons.

Duarte, N. F., Raković, M., and Santos-Victor, J. (2021). "Learning motor resonance in human-human and human-robot interaction with coupled dynamical systems," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (Xi'an, China: IEEE), 3662–3668.

Dumas, G., and Fairhurst, M. T. (2021). Reciprocity and alignment: quantifying coupling in dynamic interactions. *R. Soc. Open Sci.* 8, 210138. doi: 10.1098/rsos.210138

Dumas, G., Moreau, Q., Tognoli, E., and Kelso, J. S. (2020). The human dynamic clamp reveals the fronto-parietal network linking real-time social coordination and cognition. *Cereb. Cortex* 30, 3271–3285. doi: 10.1093/cercor/bhz308

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., and Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE* 5, e12166. doi: 10.1371/journal.pone.0012166

Ehrenreich, B. (2009). *Bright-Sided: How Positive Thinking is Undermined America*. New York, NY: Metropolitan Books.

Endo, G., Morimoto, J., Matsubara, T., Nakanishi, J., and Cheng, G. (2008). Learning CPG-based biped locomotion with a policy gradient method: application to a humanoid robot. *Int. J. Rob. Res.* 27, 213–228. doi: 10.1177/0278364907084980

Engert, V., Linz, R., and Grant, J. A. (2019). Embodied stress: the physiological resonance of psychosocial stress. *Psychoneuroendocrinology* 105, 138–146. doi: 10.1016/j.psyneuen.2018.12.221

Farber, J. (1994). Aesthetic resonance: Beyond the sign in literature. *Reader* 32, 16.

Fauché, P., Kosionis, S. G., and Lalanne, P. (2017). Collective scattering in hybrid nanostructures with many atomic oscillators coupled to an electromagnetic resonance. *Phys. Rev. B* 95, 195418. doi: 10.1103/PhysRevB.95.195418

Feldman, R. (2012). Bio-behavioral synchrony: a model for integrating biological and microsocial behavioral processes in the study of parenting. *Parenting* 12, 154–164. doi: 10.1080/15295192.2012.683342

Feldman, R. (2017). The neurobiology of human attachments. *Trends Cogn. Sci.* 21, 80–99. doi: 10.1016/j.tics.2016.11.007

Ferreira, D., and Vale, M. (2020). Geography in the big data age: an overview of the historical resonance of current debates. *Geograph. Rev.* 122, 1–17. doi: 10.1080/00167428.2020.1832424

Forlizzi, J. (2007). "How robotic products become social products: an ethnographic study of cleaning in the home," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Arlington, VA: IEEE), 129–136.

Freeman, L. M., Pang, L., and Fainman, Y. (2014). Maximizing the electromagnetic and chemical resonances of surface-enhanced Raman scattering for nucleic acids. *ACS Nano* 8, 8383–8391. doi: 10.1021/nn5028664

Fujiwara, K., Hoegen, R., Gratch, J., and Dunbar, N. E. (2021). Synchrony facilitates altruistic decision making for non-human avatars. *Comput. Human Behav.* 128, 107079. doi: 10.1016/j.chb.2021.107079

Fujiwara, K., and Yokomitsu, K. (2021). Video-based tracking approach for nonverbal synchrony: a comparison of Motion Energy Analysis and OpenPose. *Behav. Res. Methods* 53, 2700–2711. doi: 10.3758/s13428-021-01612-7

Gallese, V. (2009). Mirror neurons, embodied simulation, and the neural basis of social identification. *Psychoanal. Dialogues* 19, 519–536. doi: 10.1080/10481880903231910

Gallese, V., and Sinigaglia, C. (2018). "Embodied resonance," in *The Oxford Handbook of 4e Cognition*, eds. A. Newen, L. De Bruin, and S. Gallagher (Oxford, UK: Oxford University Press), 417–432.

Gardner, T., Aglinskas, A., and Cross, E. S. (2017). Using guitar learning to probe the Action Observation Network's response to visuomotor familiarity. *Neuroimage* 156, 174–189. doi: 10.1016/j.neuroimage.2017.04.060

Garrett, C. (1972). Tidal resonance in the Bay of Fundy and Gulf of Maine. *Nature* 238, 441–443. doi: 10.1038/238441a0

Gelfand, M. J., Caluori, N., Jackson, J. C., and Taylor, M. K. (2020). The cultural evolutionary trade-off of ritualistic synchrony. *Philos. Trans. R. Soc. B* 375, 20190432. doi: 10.1098/rstb.2019.0432

Gibbs, A. (2019). "Put a spell on you: Poetry, politics, and affective resonance in the age of the algorithm," in *Public Spheres of Resonance*, eds. A. Fleig and C. von Scheve (New York, NY: Routledge), 117–134.

Gill, S. P. (2012). Rhythmic synchrony and mediated interaction: towards a framework of rhythm in embodied interaction. *AI Soc.* 27, 111–127. doi: 10.1007/s00146-011-0362-2

Giorgi, S. (2017). The mind and heart of resonance: the role of cognition and emotions in frame effectiveness. *J. Manage. Stud.* 54, 711–738. doi: 10.1111/joms.12278

Grafsgaard, J., Duran, N., Randall, A., Tao, C., and D'Mello, S. (2018). "Generative multimodal models of nonverbal synchrony in close relationships," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an: IEEE), 195–202.

Grammer, K., Kruck, K. B., and Magnusson, M. S. (1998). The courtship dance: patterns of nonverbal synchronization in opposite-sex encounters. *J. Nonverbal Behav.* 22, 3–29. doi: 10.1023/A:1022986608835

Gratch, J., Kang, S. H., and Wang, N. (2013). Using social agents to explore theories of rapport and emotional resonance. *Soc. Emotions Nat. Artifact* 181, 2568173. doi: 10.1093/acprof:oso/9780195387643.003.0012

Gratch, J., and Lucas, G. (2021). "Rapport between humans and socially interactive agents," in *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, eds. B. Lugrin, C. Pelachaud, and D. Traum (New York, NY: Association for Computing Machinery), 433–462. doi: 10.1145/3477322.3477335

Greenberg, N. S. (2019). *Resonance and behaviors toward others are predicted by perceived warmth and competence of target* (Doctoral dissertation). Waltham, MA: Brandeis University.

Greenfield, M. D., Honing, H., Kotz, S. A., and Ravignani, A. (2021). Synchrony and rhythm interaction: from the brain to behavioural ecology. *Philos. Trans. R. Soc. B* 376, 20200324. doi: 10.1098/rstb.2020.0324

Grimes, C. A., Mungle, C. S., Zeng, K., Jain, M. K., Dreschel, W. R., Paulose, M., et al. (2002). Wireless magnetoelastic resonance sensors: a critical review. *Sensors* 2, 294–313. doi: 10.3390/s20700294

Grossberg, S. (2013). Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw.* 37, 1–47. doi: 10.1016/j.neunet.2012.09.017

Grossberg, S. (2017). Towards solving the hard problem of consciousness: the varieties of brain resonances and the conscious experiences that they support. *Neural Netw.* 87, 38–95. doi: 10.1016/j.neunet.2016.11.003

Guertin, P. A. (2013). Central pattern generator for locomotion: anatomical, physiological, and pathophysiological considerations. *Front. Neurol.* 3, 183. doi: 10.3389/fneur.2012.00183

Guevara Erra, R., Perez Velazquez, J. L., and Rosenblum, M. (2017). Neural synchronization from the perspective of non-linear dynamics. *Front. Comput. Neurosci.* 11, 98. doi: 10.3389/fncom.2017.00098

Gutsell, J. N., and Inzlicht, M. (2010). Empathy constrained: prejudice predicts reduced mental simulation of actions during observation of outgroups. *J. Exp. Soc. Psychol.* 46, 841–845. doi: 10.1016/j.jesp.2010.03.011

Gutterman, E. (2015). Easier done than said: transnational bribery, norm resonance, and the origins of the US Foreign Corrupt Practices Act. *Foreign Policy Anal.* 11, 109–128. doi: 10.1111/fpa.12027

Hall, J. A., and Schwartz, R. (2019). Empathy present and future. *J. Soc. Psychol.* 159, 225–243. doi: 10.1080/00224545.2018.1477442

Han, F., Caporale, N., and Dan, Y. (2008). Reverberation of recent visual experience in spontaneous cortical waves. *Neuron* 60, 321–327. doi: 10.1016/j.neuron.2008.08.026

Hanslmayr, S., Axmacher, N., and Inman, C. S. (2019). Modulating human memory via entrainment of brain oscillations. *Trends Neurosci.* 42, 485–499.

Hanson, L. (2009). Second-stage labor care: challenges in spontaneous bearing down. *J. Perinat. Neonatal Nurs.* 23, 31–39. doi: 10.1097/JPN.0b013e318196526b

Hanuschkin, A., Ganguli, S., and Hahnloser, R. (2013). A Hebbian learning rule gives rise to mirror neurons and links them to control theoretic inverse models. *Front. Neural Circuits* 7, 106. doi: 10.3389/fncir.2013.00106

Hasegawa, A., Okanoya, K., Hasegawa, T., and Seki, Y. (2011). Rhythmic synchronization tapping to an audio-visual metronome in budgerigars. *Sci. Rep.* 1, 1–8. doi: 10.1038/srep00120

Hasson, U., and Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philos. Trans. R. Soc. B* 371, 20150366. doi: 10.1098/rstb.2015.0366

Hattori, Y., and Tomonaga, M. (2020). Rhythmic swaying induced by sound in chimpanzees (Pan troglodytes). *Proc. Nat. Acad. Sci. U.S.A.* 117, 936–942. doi: 10.1073/pnas.1910318116

Hebb, D. O. (2005). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Psychology Press.

Hekkert, P. (2014). "Aesthetic responses to design: a battle of impulses," in *The Cambridge Handbook of the Psychology of Aesthetics and the Arts,* eds P. P. L. Tinio and J. K. Smith (Cambridge, UK: Cambridge University Press), 277–299.

Hekkert, P., Snelders, D., and Van Wieringen, P. C. (2003). 'Most advanced, yet acceptable': typicality and novelty as joint predictors of aesthetic preference in industrial design. *Br. J. Psychol.* 94, 111–124. doi: 10.1348/000712603762842147

Helfrich, R. F., Breska, A., and Knight, R. T. (2019). Neural entrainment and network resonance in support of top-down guided attention. *Curr. Opin. Psychol.* 29, 82–89. doi: 10.1016/j.copsyc.2018.12.016

Helmholtz, H. L. (2009). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge, UK: Cambridge University Press.

Henschel, A., and Cross, E. S. (2020). No evidence for enhanced likeability and social motivation towards robots after synchrony experience. *Interact. Stud.* 21, 7–23. doi: 10.1075/is.19004.hen

Henschel, A., Hortensius, R., and Cross, E. S. (2020). Social cognition in the age of human-robot interaction. *Trends Neurosci.* 43, 373–384. doi: 10.1016/j.tins.2020.03.013

Hernandez-Lallement, J., Gómez-Sotres, P., and Carrillo, M. (2020). Towards a unified theory of emotional contagion in rodents-A meta-analysis. *Neurosci. Biobehav. Rev.* 132, 1229–1248. doi: 10.1016/j.neubiorev.2020.09.010

Herrmann, C. S. (2001). Human EEG responses to 1-100 Hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Exp. Brain Res.* 137, 346–353. doi: 10.1007/s002210100682

Hicks, E., and Hicks, J. (2006). *The Law of Attraction: The Basics of the Teachings of Abraham*. London, UK: Hay House, Inc.

Himberg, T., Laroche, J., Bigé, R., Buchkowski, M., and Bachrach, A. (2018). Coordinated interpersonal behaviour in collective dance improvisation: the aesthetics of kinaesthetic togetherness. *Behav. Sci.* 8, 23. doi: 10.3390/bs8020023

Hirao, M., and Ogi, H. (1997). Electromagnetic acoustic resonance and materials characterization. *Ultrasonics* 35, 413–421. doi: 10.1016/S0041-624X(97)00030-9

Hoffman, G. (2019). Evaluating fluency in human-robot collaboration. *IEEE Trans. Human Mach. Syst.* 49, 209–218. doi: 10.1109/THMS.2019.2904558

Hoffman, G., and Ju, W. (2014). Designing robots with movement in mind. *J. Human Robot Interact.* 3, 91–122. doi: 10.5898/JHRI.3.1.Hoffman

Hoffman, G., and Weinberg, G. (2010). "Shimon: an interactive improvisational robotic marimba player," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, GA), 3097–3102.

Hofree, G., Ruvolo, P., Bartlett, M. S., and Winkielman, P. (2014). Bridging the mechanical and the human mind: spontaneous mimicry of a physically present android. *PLoS ONE* 9, e99934. doi: 10.1371/journal.pone.0099934

Hofree, G., Ruvolo, P., Reinert, A., Bartlett, M. S., and Winkielman, P. (2018). Behind the robot's smiles and frowns: in social context, people do not mirror android's expressions but react to their informational value. *Front. Neurorobot.* 12, 14. doi: 10.3389/fnbot.2018.00014

Holyoak, K. J., and Stamenković, D. (2018). Metaphor comprehension: a critical review of theories and evidence. *Psychol. Bull.* 144, 641. doi: 10.1037/bul0000145

Hore, P. J. (2015). *Nuclear Magnetic Resonance*. Oxford, UK: Oxford University Press.

Hove, M. J., and Risen, J. L. (2009). It's all in the timing: interpersonal synchrony increases affiliation. *Soc. Cogn.* 27, 949–960. doi: 10.1521/soco.2009.27.6.949

Howie, P. C., and Bagnall, R. G. (2020). A criterial framework for concept evaluation in social practice and scholarship. *Sociol. Methods Res.* 49, 719–741. doi: 10.1177/0049124118769104

Hughes, J. (2012). "Compassionate AI and selfless robots: a buddhist approach," *Robot Ethics: The Ethical and Social Implications of Robotics* eds. P. Lin, K. Abney, and G. A. Bekey (Cambridge, MA: MIT Press), 69–83.

Hummels, C. C. M., Ross, P. R., and Overbeeke, C. J. (2003). "In search of resonant human computer interaction: building and testing aesthetic installations," in *Proceedings of the 9th IFIP Conference on Human-Computer Interaction* (Zurich), 399–406.

Hunt, T., and Schooler, J. W. (2019). The easy part of the hard problem: a resonance theory of consciousness. *Front. Hum. Neurosci.* 13, 378. doi: 10.3389/fnhum.2019.00378

Hurley, S. (2008). The shared circuits model (SCM): how control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behav. Brain Sci.* 31, 1–22. doi: 10.1017/S0140525X07003123

Hutchins, E. (2001). "Distributed cognition," in *International Encyclopedia of the Social and Behavioral Sciences*, eds. N. J. Smelser and P. B. Baltes (Amsterdam, NL: Elsevier Science), 138.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annu. Rev. Psychol.* 60, 653–670. doi: 10.1146/annurev.psych.60.110707.163604

Iio, T., Shiomi, M., Shinozawa, K., Akimoto, T., Shimohara, K., and Hagita, N. (2011). Investigating entrainment of people's pointing gestures by robot's gestures using a WOZ method. *Int. J. Soc. Robot.* 3, 405–414. doi: 10.1007/s12369-011-0112-0

Iio, T., Shiomi, M., Shinozawa, K., Shimohara, K., Miki, M., and Hagita, N. (2015). Lexical entrainment in human robot interaction. *Int. J. Soc. Robot.* 7, 253–263. doi: 10.1007/s12369-014-0255-x

Ijspeert, A. J. (2008). Central pattern generators for locomotion control in animals and robots: a review. *Neural Netw.* 21, 642–653. doi: 10.1016/j.neunet.2008.03.014

Iqbal, T., and Riek, L. D. (2021). "Temporal anticipation and adaptation methods for fluent human-robot teaming," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 3736–3743.

Iversen, J. R., Patel, A. D., Nicodemus, B., and Emmorey, K. (2015). Synchronization to auditory and visual rhythms in hearing and deaf individuals. *Cognition* 134, 232–244. doi: 10.1016/j.cognition.2014.10.018

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience*. Cambridge, MA: MIT Press.

Jackson, S. A., and Csikszentmihalyi, M. (1999). *Flow in Sports*. Champaign, IL: Human Kinetics.

Jiang, L. (2020, October 22). Reasons the largest chinese restaurant in the netherlands uses the pudu robotics food delivery robot. *DutchNews.nl.* Available online at: https://www.dutchnews.nl/news/businesswire/reasons-the-largest-chinese-restaurant-in-the-netherlands-uses-the-pudu-robotics-food-delivery-robot-2/

Jouaiti, M., and Henaff, P. (2018). "CPG-based controllers can trigger the emergence of social synchrony in human-robot interactions," in *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (Genoa, Italy: IEEE), 71–76.

Kaneshiro, B., Nguyen, D. T., Norcia, A. M., Dmochowski, J. P., and Berger, J. (2020). Natural music evokes correlated EEG responses reflecting temporal structure and beat. *Neuroimage* 214, 116559. doi: 10.1016/j.neuroimage.2020.116559

Kaneshiro, B., Nguyen, D. T., Norcia, A. M., Dmochowski, J. P., and Berger, J. (2021). Inter-subject EEG correlation reflects time-varying engagement with natural music. *bioRxiv*. 2021, 1–32. doi: 10.1101/2021.04.14.439913

Katz, B. F. (1995). Harmonic resolution, neural resonance, and positive affect. *Music Percept.* 13, 79–108. doi: 10.2307/40285686

Keller, K. L. (2010). Brand equity management in a multichannel, multimedia retail environment. *J. Interact. Market.* 24, 58–70. doi: 10.1016/j.intmar.2010.03.001

Kendon, A. (1970). Movement coordination in social interaction: some examples described. *Acta Psychol.* 32, 101–125. doi: 10.1016/0001-6918(70)90094-6

Keysers, C., and Gazzola, V. (2009). Expanding the mirror: vicarious activity for actions, emotions, and sensations. *Curr. Opin. Neurobiol.* 19, 666–671. doi: 10.1016/j.conb.2009.10.006

Keysers, C., and Gazzola, V. (2014). Hebbian learning and predictive mirror neurons for actions, sensations and emotions. *Philos. Trans. R. Soc. B* 369, 20130175. doi: 10.1098/rstb.2013.0175

Kingsbury, L., and Hong, W. (2020). A multi-brain framework for social interaction. *Trends Neurosci.* 43, 651–666. doi: 10.1016/j.tins.2020.06.008

Kinreich, S., Djalovski, A., Kraus, L., Louzoun, Y., and Feldman, R. (2017). Brain-to-brain synchrony during naturalistic social interactions. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-17339-5

Kirkwood, G. L., Otmar, C. D., and Hansia, M. (2021). Who's leading this dance?: Theorizing automatic and strategic synchrony in human-exoskeleton interactions. *Front. Psychol.* 12, 624108. doi: 10.3389/fpsyg.2021.624108

Kirsch, L. P., Snagg, A., Heerey, E., and Cross, E. S. (2016a). The impact of experience on affective responses during action observation. *PLoS ONE* 11, e0154681. doi: 10.1371/journal.pone.0154681

Kirsch, L. P., Urgesi, C., and Cross, E. S. (2016b). Shaping and reshaping the aesthetic brain: Emerging perspectives on the neurobiology of embodied aesthetics. *Neurosci. Biobehav. Rev.* 62, 56–68. doi: 10.1016/j.neubiorev.2015.12.005

Kirschner, S., and Tomasello, M. (2009). Joint drumming: social context facilitates synchronization in preschool children. *J. Exp. Child Psychol.* 102, 299–314. doi: 10.1016/j.jecp.2008.07.005

Klimesch, W. (2018). The frequency architecture of brain and brain body oscillations: an analysis. *Euro. J. Neurosci.* 48, 2431–2453. doi: 10.1111/ejn.14192

Knight, R. T., and Nakada, T. (1998). Cortico-limbic circuits and novelty: a review of EEG and blood flow data. *Rev. Neurosci.* 9, 57–70. doi: 10.1515/REVNEURO.1998.9.1.57

Koban, L., Ramamoorthy, A., and Konvalinka, I. (2019). Why do we fall into sync with others? Interpersonal synchronization and the brain's optimization principle. *Soc. Neurosci.* 14, 1–9. doi: 10.1080/17470919.2017.1400463

Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Commun.* 52, 587–597. doi: 10.1016/j.specom.2010.02.007

Kozima, H., Michalowski, M. P., and Nakagawa, C. (2009). Keepon. *Int. J. Soc. Robot.* 1, 3–18. doi: 10.1007/s12369-008-0009-8

Kühne, V., Rosenthal-von der Pütten, A. M., and Krämer, N. C. (2013). "Using linguistic alignment to enhance learning experience with pedagogical agents: the special case of dialect," in *International Workshop on Intelligent Virtual Agents*, eds. R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira (Berlin: Springer), 149–158.

Kumagai, Y., Matsui, R., and Tanaka, T. (2018). Music familiarity affects EEG entrainment when little attention is paid. *Front. Hum. Neurosci.* 12, 444. doi: 10.3389/fnhum.2018.00444

Lakatos, P., Joachim, G., and Gregor, T. (2019). A new unifying account of the roles of neuronal entrainment. *Curr. Biol.* 29, R890–R905. doi: 10.1016/j.cub.2019.07.075

Lakoff, G., and Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cogn. Sci.* 4, 195–208. doi: 10.1207/s15516709cog0402_4

Lampl, I., and Yarom, Y. (1997). Subthreshold oscillations and resonant behavior: two manifestations of the same mechanism. *Neuroscience* 78, 325–341. doi: 10.1016/S0306-4522(96)00588-X

Large, E. W. (2010). "Neurodynamics of music," in *Music Perception*, eds. M. R. Jones, R. R. Fay, A. N. Popper (New York, NY: Springer), 201–231.

Large, E. W., and Snyder, J. S. (2009). Pulse and meter as neural resonance. *Ann. N. Y. Acad. Sci.* 1169, 46–57. doi: 10.1111/j.1749-6632.2009.04550.x

Larsson, M., Richter, J., and Ravignani, A. (2019). Bipedal steps in the development of rhythmic behavior in humans. *Music Sci.* 2, 2059204319892617. doi: 10.1177/2059204319892617

Lassiter, J. B., Sobhani, H., Fan, J. A., Kundu, J., Capasso, F., Nordlander, P., et al. (2010). Fano resonances in plasmonic nanoclusters: geometrical and chemical tunability. *Nano Lett.* 10, 3184–3189. doi: 10.1021/nl102108u

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., et al. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions in married couples, in *Proceedings of InterSpeech* (Makuhari: International Speech Communication Association), 793–796.

Lee, C. S., Kao, Y. F., Kuo, Y. H., and Wang, M. H. (2007). Automated ontology construction for unstructured text documents. *Data Knowl. Eng.* 60, 547–566. doi: 10.1016/j.datak.2006.04.001

Lee, M. K., Forlizzi, J., Rybski, P. E., Crabbe, F., Chung, W., Finkle, J., Glaser, E., and Kiesler, S. (2009). "The snackbot: documenting the design of a robot for long-term human-robot interaction," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, CA), 7–14.

Lehar, S. (2003). Harmonic resonance theory: an alternative to the "Neuron Doctrine" paradigm of neurocomputation to address gestalt properties of perception. *Behav. Brain Sci.* 26, 15. doi: 10.1017/S0140525X03000098

Levantino, V. (2018). *BAO-ME: design of a huggable device to improve mental well-being through haptic interactions* (Keio University Master's Thesis). Minato, Tokyo, Japan: Sugiura, Kazunori 杉浦,一Publisher 塾大学大学院メディアデザイン研究科.

Lewis, T., Amini, F., and Lannon, R. (2001). *A General Theory of Love*. New York, NY: Vintage.

Li, J., Liu, Q., Zhai, Y., Molinas, M., and Wu, M. (2020). Analysis of harmonic resonance for locomotive and traction network interacted system considering the frequency-domain passivity properties of the digitally controlled converter. *Front. Energy Res.* 276, 1–16. doi: 10.3389/fenrg.2020.523333

Lim, V., Rooksby, M., and Cross, E. S. (2021). Social robots on a global stage: establishing a role for culture during human-robot interaction. *Int. J. Soc. Robot*. 13, 1307–1333. doi: 10.1007/s12369-020-00710-4

Lin, A., and Lomas, J. D. (2022). "Enigma of mind: a theory of evolution and conscious experience," in *Enigma: Darwin College Lecture Series*, eds. E. Ward and J. Gibson (Cambridge: Cambridge University Press).

Lipps, T. (1891). Zweiter ästhetischer Literatturbericht. *Philosophische Monatshefte* 27, 161–182.

Lo, S. Y., and Huang, H. P. (2016). Realization of sign language motion using a dual-arm/hand humanoid robot. *Intelligent Service Robot.* 9, 333–345. doi: 10.1007/s11370-016-0203-8

Lobis, S. (2015). *The Virtue of Sympathy: Magic, Philosophy, and Literature in Seventeenth-Century England* (New Haven, CT: Yale University Press).

Lolo and Nate (2022). *17 Weird to Wonderful Long Distance Relationship Gadgets.* Available online at: https://lastingthedistance.com/long-distance-relationship-gadgets/

Lomas, J. D., Forlizzi, J., Poonwala, N., Patel, N., Shodhan, S., Patel, K., et al. (2016). "Interface design optimization as a multi-armed bandit problem," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA), 4142–4153.

Lomas, J. D., Karac, M., and Gielen, M. (2021). Design space cards: using a card deck to navigate the design space of interactive play. *Proc. ACM Human Computer Interact.* 5, 227. doi: 10.1145/3474654

Lomas, J. D., and Xue, H. (2022). Harmony in design: A synthesis of literature from classical philosophy, the sciences, economics, and design. *She Ji: The Journal of Design, Economics, and Innovation* 8, 5–64.

Lorenz, T., Weiss, A., and Hirche, S. (2016). Synchrony and reciprocity: key mechanisms for social companion robots in therapy and care. *Int. J. Soc. Robot*. 8, 125–143. doi: 10.1007/s12369-015-0325-8

Ludeke, C. A. (1942). Resonance. *J. Appl. Phys.* 13, 418–423. doi: 10.1063/1.1714888

Lupetti, M. L., and Van Mechelen, M. (2022). "Promoting children's critical thinking towards robotics through robot deception," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan), 588–597.

MacDougall, H. G., and Moore, S. T. (2005). Marching to the beat of the same drummer: the spontaneous tempo of human locomotion. *J. Appl. Physiol.* 99, 1164–1173. doi: 10.1152/japplphysiol.00138.2005

Macik, M., Prazakova, K., Kutikova, A., Mikovec, Z., Adolf, J., Havlik, J., et al. (2017). "Breathing friend: tackling stress through portable tangible breathing artifact," in *IFIP Conference on Human-Computer Interaction* (Cham: Springer), 106–115.

Madsen, J., Margulis, E. H., Simchy-Gross, R., and Parra, L. C. (2019). Music synchronizes brainwaves across listeners with strong effects of repetition, familiarity and training. *Sci. Rep.* 9, 1–8. doi: 10.1038/s41598-019-40254-w

Matheny, T., Poe, P., Fisher, M., and Warren, S. (2018). Chaos and stability in Donald Trump's acceptance speech. *Media Watch* 9, 260–266. doi: 10.15655/mw/2018/v9i3/49499

Mayo, O., and Gordon, I. (2020). In and out of synchrony-Behavioral and physiological dynamics of dyadic interpersonal coordination. *Psychophysiology* 57, e13574. doi: 10.1111/psyp.13574

McDonnell, T. E., Bail, C. A., and Tavory, I. (2017). A theory of resonance. *Sociol. Theory* 35, 1–14. doi: 10.1177/0735275117692837

McGarva, A. R., and Warner, R. M. (2003). Attraction and social coordination: mutual entrainment of vocal activity rhythms. *J. Psycholinguist. Res.* 32, 335–354. doi: 10.1023/A:1023547703110

McMillan, W. L., and Anderson, P. W. (1966). Theory of geometrical resonances in the tunneling characteristics of thick films of superconductors. *Phys. Rev. Lett.* 16, 85. doi: 10.1103/PhysRevLett.16.85

McQuarrie, E. F., and Mick, D. G. (1992). On resonance: a critical pluralistic inquiry into advertising rhetoric. *J. Consumer Res.* 19, 180–197. doi: 10.1086/209295

Melnyk, A., and Hénaff, P. (2019). Physical analysis of handshaking between humans: mutual synchronisation and social context. *Int. J. Soc. Robot*. 11, 541–554. doi: 10.1007/s12369-019-00525-y

Metzler, T. A., Lewis, L. M., and Pope, L. C. (2016). Could robots become authentic companions in nursing care?. *Nurs. Philos.* 17, 36–48. doi: 10.1111/nup.12101

Mintzberg, H., Quinn, J. B., and Voyer, J. (1995). *The Strategy Process*. Collegiate edition. Englewood Cliffs, NJ: Prentice-Hall.

Mohammadi-Khanaman, S., and Lundström, C. (2019). *Somnox Care Robot: Development of a Soft Robot for Improved Sleep*. Chalmers University of Technology Repository. Available online at: https://hdl.handle.net/20.500.12380/300447

Molkov, Y. I., Zoccal, D. B., Baekey, D. M., Abdala, A. P., Machado, B. H., Dick, T. E., et al. (2014). Physiological and pathophysiological interactions between the respiratory central pattern generator and the sympathetic nervous system. *Prog. Brain Res.* 212, 1–23. doi: 10.1016/B978-0-444-63488-7.00001-X

Moran, N., Pellegrino, D., Slingerland, J. K., and Kells, G. (2017). Parafermionic clock models and quantum resonance. *Phys. Rev. B* 95, 235127. doi: 10.1103/PhysRevB.95.235127

Moreau, Q., and Dumas, G. (2021). Beyond "correlation vs. causation": multi-brain neuroscience needs explanation. *Trends Cogn. Sci.* 25, 542–543. doi: 10.1016/j.tics.2021.02.011

Morgan, R., Fischer, R., and Bulbulia, J. A. (2017). To be in synchrony or not? A meta-analysis of synchrony's effects on behavior, perception, cognition and affect. *J. Exp. Soc. Psychol.* 72, 13–20. doi: 10.1016/j.jesp.2017.03.009

Mühlhoff, R. (2019). "Affective resonance," in *Affective Societies* (Routledge), 189–199.

Mutlu, B., Roy, N., and Šabanović, S. (2016). "Cognitive human-robot interaction," in *Springer Handbook of Robotics*, eds. B. Siciliano and O. Khatib (Würzburg, Germany: Springer), 1907–1934.

New York Times (2021, December 9) *Will Smith is Done Trying to Be Perfect*. Available online at: https://www.nytimes.com/2021/12/09/magazine/will-smith-interview.html

Northoff, G., Wainio-Theberge, S., and Evers, K. (2020). Is temporo-spatial dynamics the "common currency" of brain and mind? In Quest of "Spatiotemporal Neuroscience". *Phys. Life Rev.* 33, 34–54. doi: 10.1016/j.plrev.2019.05.002

Oldenburg, S. J., Averitt, R. D., Westcott, S. L., and Halas, N. J. (1998). Nanoengineering of optical resonances. *Chem. Phys. Lett.* 288, 243–247. doi: 10.1016/S0009-2614(98)00277-2

O'Loughlin, H. (2018). *Effect of a heart beating breathing doll on infant sleep* (Master's Thesis). University of Canterbury. Available online at: http://hdl.handle.net/10092/15580

Orgs, G., Dombrowski, J., Heil, M., and Jansen-Osmann. (2008). Expertise in dance modulates alpha/beta event-related desynchronization during action observation. *Euro. J. Neurosci.* 27, 3380–3384. doi: 10.1111/j.1460-9568.2008.06271.x

Paasonen, S. (2011). *Carnal Resonance: Affect and Online Pornography*. Cambridge, MA: MIT Press.

Paladino, M.-P., Mazzurega, M., Pavani, F., and Schubert, T. W. (2010). Synchronous multisensory stimulation blurs self-other boundaries. *Psychol. Sci.* 21, 1202–1207. doi: 10.1177/0956797610379234

Pandey, P., Ahmad, N., Miyapuram, K. P., and Lomas, D. (2021). "Predicting dominant beat frequency from brain responses while listening to music," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Houston, TX: IEEE), 3058–3064.

Patel, A., Iversen, J., Bregman, M., and Schulz, I. (2009). Studying synchronization to a musical beat in nonhuman animals. *Ann. N. Y. Acad. Sci.* 1169, 459–469. doi: 10.1111/j.1749-6632.2009.04581.x

Patel, A. D., and Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: the Action Simulation for Auditory Prediction (ASAP) hypothesis. *Front. Syst. Neurosci.* 8, 57. doi: 10.3389/fnsys.2014.00057

Pérez, P., Madsen, J., Banellis, L., Türker, B., Raimondo, F., Perlbarg, V., et al. (2021). Conscious processing of narrative stimuli synchronizes heart rate between individuals. *Cell Rep.* 36, 109692. doi: 10.1016/j.celrep.2021.109692

Pikovsky, A., Rosenblum, M., and Kurths, J. (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge, UK: Cambridge University Press.

Porter, M. E. (1991). Towards a dynamic theory of strategy. *Strategic Manage. J.* 12, 95–117. doi: 10.1002/smj.4250121008

Poshakinskiy, A. V., and Astakhov, G. V. (2019). Optically detected spin-mechanical resonance in silicon carbide membranes. *Phys. Rev. B* 100, 094104. doi: 10.1103/PhysRevB.100.094104

Prasad, L. (2019). Ethical resonance: the concept, the practice, and the narration. *J. Religious Ethics* 47, 394–415. doi: 10.1111/jore.12261

Preston, S. D., and De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behav. Brain Sci.* 25, 1–20. doi: 10.1017/S0140525X02000018

Prochazkova, E., Sjak-Shie, E., Behrens, F., Lindh, D., and Kret, M. E. (2022). Physiological synchrony is associated with attraction in a blind date setting. *Nat. Human Behav.* 6, 269–278. doi: 10.1038/s41562-021-01197-3

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress Neurobiol.* 160, 1–44. doi: 10.1016/j.pneurobio.2017.07.001

Quach, J. Q. (2016). Spin gravitational resonance and graviton detection. *Phys. Rev. D* 93, 104048. doi: 10.1103/PhysRevD.93.104048

Raja, V. (2018). *A story of resonance* (Doctoral Dissertation). Available online at: https://etd.ohiolink.edu/!etd.send_file?accession=ucin1530272507861696&disposition=inline

Raja, V. (2020). Resonance and radical embodiment. *Synthese*. 199 1–29. doi: 10.1007/s11229-020-02610-6

Rajasekar, S., and Sanjuan, M. A. (2016). *Nonlinear Resonances*. Switzerland: Springer International Publishing.

Reinero, D. A., Dikker, S., and Van Bavel, J. J. (2021). Inter-brain synchrony in teams predicts collective performance. *Soc. Cogn. Affect. Neurosci.* 16, 43–57. doi: 10.1093/scan/nsaa135

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., and Völlm, B. A. (2011). The QCAE: a questionnaire of cognitive and affective empathy. *J. Pers. Assess.* 93, 84–95. doi: 10.1080/00223891.2010.528484

Repp, B. H., and Su, Y. H. (2013). Sensorimotor synchronization: a review of recent research (2006-2012). *Psychon. Bull. Rev.* 20, 403–452. doi: 10.3758/s13423-012-0371-2

Roger, P. (2020). Toward a theory of resonance in cinema-On Claudia von Alemann's Le Voyage à Lyon. *Écrans* 2019, 65–78. doi: 10.15122/isbn.978-2-406-10499-5.p.0065

Roll, G., and Schweiger, G. (2000). Geometrical optics model of Mie resonances. *JOSA A* 17, 1301–1311. doi: 10.1364/JOSAA.17.001301

Ronsse, R., Vitiello, N., Lenzi, T., Van Den Kieboom, J., Carrozza, M. C., and Ijspeert, A. J. (2010). Human-robot synchrony: flexible assistance using adaptive oscillators. *IEEE Trans. Biomed. Eng.* 58, 1001–1012. doi: 10.1109/TBME.2010.2089629

Rosa, H. (2018). The idea of resonance as a sociological concept. *Global Dial.* 8, 41–44. Available online at: https://globaldialogue.isa-sociology.org/articles/the-idea-of-resonance-as-a-sociological-concept (accessed April 6, 2022)

Rosa, H. (2019). *Resonance: A Sociology of Our Relationship to the World*. Cambridge, UK: John Wiley & Sons.

Ruthven, I. (2021). Resonance and the experience of relevance. *J. Assoc. Information Sci. Technol.* 72, 554–569. doi: 10.1002/asi.24424

Sadler, P., Ethier, N., Gunn, G. R., Duong, D., and Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *J. Personal. Soc. Psychol.* 97, 1005. doi: 10.1037/a0016232

Safron, A. (2016). What is orgasm? A model of sexual trance and climax via rhythmic entrainment. *Socioaffective Neurosci. Psychol.* 6, 31763. doi: 10.3402/snp.v6.31763

Safron, A. (2020). An Integrated World Modeling Theory (IWMT) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. *Front. Artificial Intell.* 3, 30. doi: 10.3389/frai.2020.00030

Santos, M., and Egerstedt, M. (2021). From motions to emotions: can the fundamental emotions be expressed in a robot swarm?. *Int. J. Soc. Robot*. 13, 751–764. doi: 10.1007/s12369-020-00665-6

Sarma, G. P., Hay, N. J., and Safron, A. (2018). "AI safety and reproducibility: establishing robust foundations for the neuropsychology of human values, in *International Conference on Computer Safety, Reliability, and Security* (Cham: Springer), 507–512.

Sato, T. G., and Moriya, T. (2019). "Respiration rate change induced by controlling the phasic relationship between melodic sound and respiration," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin, Germany: IEEE), 3123–3126.

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., et al. (2021). Music as a coevolved system for social bonding. *Behav. Brain Sci.* 44, 1–36. doi: 10.1017/S0140525X20000333

Schachner, A., Brady, T. F., Pepperberg, I. M., and Hauser, M. D. (2009). Spontaneous motor entrainment to music in multiple vocal mimicking species. *Curr. Biol.* 19, 831–836. doi: 10.1016/j.cub.2009.03.061

Schemer, C., Wirth, W., and Matthes, J. (2012). Value resonance and value framing effects on voting intentions in direct-democratic campaigns. *Am. Behav. Scientist* 56, 334–352. doi: 10.1177/00027642114 26329

Schliesser, E. (ed.). (2015). *Sympathy: A history*. Oxford, UK: Oxford University Press.

Schrock, D., Holden, D., and Reid, L. (2004). Creating emotional resonance: interpersonal emotion work and motivational framing in a transgender community. *Soc. Probl.* 51, 61–81. doi: 10.1525/sp.2004. 51.1.61

Schubert, E., Vincs, K., and Stevens, C. J. (2013). Identifying regions of good agreement among responders in engagement with a piece of live dance. *Empirical Stud. Arts* 31, 1–20. doi: 10.2190/EM.31.1.a

Schütz-Bosbach, S., and Prinz, W. (2007). Perceptual resonance: action-induced modulation of perception. *Trends Cogn. Sci.* 11, 349–355. doi: 10.1016/j.tics.2007.06.005

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., Pozzo, T., and Sandini, G. (2012). Measuring human-robot interaction through motor resonance. *Int. J. Soc. Robot.* 4, 223–234.

Sciutti, A., and Sandini, G. (2017). Interacting with robots to investigate the bases of social interaction. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 2295–2304. doi: 10.1109/TNSRE.2017.2753879

Shang, S. S., Wu, Y. L., and Sie, Y. J. (2017). Generating consumer resonance for purchase intention on social network sites. *Comput. Human Behav.* 69, 18–28. doi: 10.1016/j.chb.2016.12.014

Sharp, H., Preece, J., and Rogers, Y. (2019). *Interaction Design: Beyond Human Computer Interaction, 5th Edn*. Chichester: Wiley.

Shaw, M. (2011). The role of design spaces. *IEEE Software* 29, 46–50. doi: 10.1109/MS.2011.121

Sheldrake, R. (2011). *The Presence of the Past: Morphic Resonance and the Habits of Nature*. London, UK: Icon Books Ltd.

Shim, S. B., Imboden, M., and Mohanty, P. (2007). Synchronized oscillation in coupled nanomechanical oscillators. *Science* 316, 95–99. doi: 10.1126/science.1137307

Shneiderman, B. (2022). *Human-Centered AI*. Oxford, UK: Oxford University Press.

Siegel, I. R. (2013). Therapist as a container for spiritual resonance and client transformation in transpersonal psychotherapy: an exploratory heuristic study. *J. Transpersonal Psychol.* 45, 49–74.

Sinclair, A. T. (1975). The orbital resonance amongst the Galilean satellites of Jupiter. *Mon. Not. R. Astron. Soc.* 171, 59–72. doi: 10.1093/mnras/ 171.1.59

Slichter, C. P. (2013). *Principles of Magnetic Resonance,* Vol. 1. London, UK: Springer Science & Business Media.

Slow Mo Guys (2021, July 15). Shattering a wine glass with sound at 187,500FPS [Video]. *Youtube*. Available online at: https://youtu.be/X3uKmTdgmQg?t=339

Smith, A. (1759). *Theory of Moral Sentiments*, eds A. Kincaid and J. Bell (Edinburgh, UK: Strand Publishing).

Snow, D. A., and Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization. *Int. Soc. Movement Res.* 1, 197–217.

Stern, D. N., Hofer, L., Haft, W., and Dore, J. (1985). "Affect attunement: the sharing of feeling states between mother and infant by means of inter-modal fluency," in *Social Perception in Infants*, eds. T. Field and N. A. Fox (New York, NK: Ablex Publishing Corporation), 249–268.

Stiefel, K. M., and Ermentrout, G. B. (2016). Neurons as oscillators. *J. Neurophysiol.* 116, 2950–2960. doi: 10.1152/jn.00525.2015

Stray, J., Vendrov, I., Nixon, J., Adler, S., and Hadfield-Menell, D. (2021). What are you optimizing for aligning recommender systems with human values. *arXiv preprint arXiv*:2107.10939. doi: 10.48550/arXiv.2107.10939

Strogatz, S. H., Abrams, D. M., McRobie, A., Eckhardt, B., and Ott, E. (2005). Crowd synchrony on the millennium bridge. *Nature* 438, 43–44. doi: 10.1038/438043a

Strogatz, S. J. P. T. (2003). Review of Synchronization (Book). *Phys Today*. 56, 47–47.

Strydom, P. (2003). "Resonance: triggering a dormant dimension of the public sphere," in *European Social Theory: Sources and Challenges*, eds V. Kultygin and V. Zhukov (European Social Theory, Moscow: European Sociological Association), 83–99.

Sundfors, R. K., Bolef, D. I., and Fedders, P. A. (1983). Nuclear acoustic resonance in metals and alloys: a review. *Hyperfine Interact.* 14, 271–313. doi: 10.1007/BF02043303

Suzuki, N., and Katagiri, Y. (2007). Prosodic alignment in human-computer interaction. *Conn. Sci.* 19, 131–141. doi: 10.1080/09540090701369125

Swann, W. B. Jr., Jetten, J., Gómez, Á., Whitehouse, H., and Bastian, B. (2012). When group membership gets personal: a theory of identity fusion. *Psychol. Rev.* 119, 441. doi: 10.1037/a0028589

Takeya, R., Kameda, M., Patel, A. D., and Tanaka, M. (2017). Predictive and tempo-flexible synchronization to a visual metronome in monkeys. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-06417-3

Tegnér, J., Compte, A., and Wang, X. J. (2002). The dynamical stability of reverberatory neural circuits. *Biol. Cybern.* 87, 471–481. doi: 10.1007/s00422-002-0363-9

Titchener, E. B. (1909). *Lectures on the Experimental Psychology of the Thought Processes*. New York, NY: The MacMillan Company.

Tojo, S., Taguchi, Y., Masuyama, Y., Hayashi, T., Saito, H., and Hirano, T. (2010). Controlling phase separation of binary Bose-Einstein condensates via mixed-spin-channel Feshbach resonance. *Phys. Rev. A* 82, 033609. doi: 10.1103/PhysRevA.82.033609

Tonks, L. (1931). Plasma-electron resonance, plasma resonance and plasma shape. *Phys. Rev.* 38, 1219. doi: 10.1103/PhysRev.38.1219

Treal, T., Jackson, P. L., Jeuvrey, J., Vignais, N., and Meugnot, A. (2021). Natural human postural oscillations enhance the empathic response to a facial pain expression in a virtual character. *Sci. Rep.* 11, 1–10. doi: 10.1038/s41598-021-91710-5

Trehub, S. E., and Trainor, L. (1998). Singing to infants: Lullabies and play songs. *Adv. Infancy Res.* 12, 43–78.

Trost, W. J., Labbé, C., and Grandjean, D. (2017). Rhythmic entrainment as a musical affect induction mechanism. *Neuropsychologia* 96, 96–110. doi: 10.1016/j.neuropsychologia.2017.01.004

Uithol, S., van Rooij, I., Bekkering, H., and Haselager, P. (2011). Understanding motor resonance. *Soc. Neurosci.* 6, 388–397. doi: 10.1080/17470919.2011.559129

Valencia, A. L., and Froese, T. (2020). What binds us? Inter-brain neural synchronization and its implications for theories of human consciousness. *Neurosci. Consciousness* 2020, niaa010. doi: 10.1093/nc/niaa010

van Werven, R., Bouwmeester, O., and Cornelissen, J. P. (2019). Pitching a business idea to investors: how new venture founders use micro-level rhetoric to achieve narrative plausibility and resonance. *Int. Small Business J.* 37, 193–214. doi: 10.1177/0266242618818249

Varga, S., and Heck, D. H. (2017). Rhythms of the body, rhythms of the brain: respiration, neural oscillations, and embodied cognition. *Conscious. Cogn.* 56, 77–90. doi: 10.1016/j.concog.2017.09.008

Vilibić, I. (2008). Numerical simulations of the Proudman resonance. *Cont. Shelf Res.* 28, 574–581. doi: 10.1016/j.csr.2007.11.005

Vincent, U. E., McClintock, P. V., Khovanov, I. A., and Rajasekar, S. (2021). Vibrational and stochastic resonances in driven nonlinear systems. *Philos. Trans. R. Soc. A* 379, 20200226. doi: 10.1098/rsta.2020.0226

Vlemincx, E., Abelson, J. L., Lehrer, P. M., Davenport, P. W., Van Diest, I., and Van den Bergh, O. (2013). Respiratory variability and sighing: a psychophysiological reset model. *Biol Psychol.* 93, 24–32.

Wang, S., Lin, D. N. C., Zheng, X., and Ji, J. (2021). Departure from the exact location of mean motion resonances induced by the gas disk in systems observed by Kepler. *Astron. J.* 161, 77. doi: 10.3847/1538-3881/abcfb9

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* 24, 455–463. doi: 10.1016/S0166-2236(00)01868-3

Warren, L. (2004). "Negotiating entrepreneurial identity: communities of practice and changing discourses." *Int. J. Entrepreneurship Innov.* 5, 25–35. doi: 10.5367/000000004772913764

Weber, R., and Fisher, J. T. (2020). "Advancing the synchronization theory of flow experiences," in *The Handbook of Communication Science and Biology*, eds. K. Floyd and R. Weber (London, UK: Routledge), 157–176.

Weber, R., Tamborini, R., Westcott-Baker, A., and Kantor, B. (2009). Theorizing flow and media enjoyment as cognitive synchronization of attentional and reward networks. *Commun. Theory* 19, 397–422. doi: 10.1111/j.1468-2885.2009.01352.x

Wertz, J. (2012). *Electron Spin Resonance: Elementary Theory and Practical Applications*. Berlin, Germany: Springer Science & Business Media.

Wheatley, T., and Sievers, B. (2016). Toward a neuroscience of social resonance. *Positive Neurosci.* 37, 37–54. doi: 10.1093/acprof:oso/9780199977925.003.0004

Wilfinger, R. J., Bardell, P. H., and Chhabra, D. S. (1968). The resonistor: a frequency selective device utilizing the mechanical resonance of a silicon substrate. *IBM J. Res. Dev.* 12, 113–118. doi: 10.1147/rd.121.0113

Williams, E. R. (1992). The Schumann resonance: a global tropical thermometer. *Science* 256, 1184–1187. doi: 10.1126/science.256.5060.1184

Wilson, M., and Cook, P. F. (2016). Rhythmic entrainment: why humans want to, fireflies can't help it, pet birds try, and sea lions have to be bribed. *Psychon. Bull. Rev.* 23, 1647–1659. doi: 10.3758/s13423-016-1013-x

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* 12, 957–968. doi: 10.3758/BF03206432

Wiltermuth, S. (2012a). Synchrony and destructive obedience. *Soc. Influence* 7, 78–89. doi: 10.1080/15534510.2012.658653

Wiltermuth, S. (2012b). Synchronous activity boosts compliance with requests to aggress. *J. Exp. Soc. Psychol.* 48, 453–456. doi: 10.1016/j.jesp.2011.10.007

Winters, R. M., Walker, B. N., and Leslie, G. (2021). "Can you hear my heartbeat?: hearing an expressive biosignal elicits empathy," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan), 1–11.

Witek, M. A. (2019). "Feeling at one: socio-affective distribution, vibe, and dance-music consciousness," in *Music and Consciousness 2: Worlds, Practices, Modalities*, eds R. Herbert, E. Clarke, and D. Clarke (New York, NY: Oxford University Press), 93–112.

Wohltjen, S., and Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *Proc. Nat. Acad. Sci. U.S.A.* 118, 1–11. doi: 10.1073/pnas.2106645118

Wolf, E. (2010). Statistical similarity as a unifying concept of the theories of coherence and polarization of light. *Opt. Commun.* 283, 4427–4429. doi: 10.1016/j.optcom.2010.04.091

Wood, A., Rychlowska, M., Korb, S., and Niedenthal, P. (2016). Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends Cogn. Sci.* 20, 227–240. doi: 10.1016/j.tics.2015.12.010

Wood, C., Caldwell-Harris, C., and Stopa, A. (2018). The rhythms of discontent: synchrony impedes performance and group functioning in an interdependent coordination task. *J. Cogn. Cult.* 18, 154–179. doi: 10.1163/15685373-12340028

Xie, F., Zuo, Q., Chen, Q., Fang, H., He, K., Du, R., et al. (2021). Designs of the biomimetic robotic fishes performing body and/or caudal fin (BCF) swimming locomotion: a review. *J. Intelligent Robotic Syst.* 102, 1–19. doi: 10.1007/s10846-021-01379-1

Yang, J. (2014). The role of the right hemisphere in metaphor comprehension: a meta-analysis of functional magnetic resonance imaging studies. *Hum. Brain Mapp.* 35, 107–122. doi: 10.1002/hbm.22160

Yang, T., Chen, J., Huang, L., Wang, T., Xiao, C., Sun, Z., et al. (2015). Extremely short-lived reaction resonances in Cl+ HD (v= 1)→ DCl+ H due to chemical bond softening. *Science* 347, 60–63. doi: 10.1126/science.1260527

Yaniv, A. U., Salomon, R., Waidergoren, S., Shimon-Raz, O., Djalovski, A., and Feldman, R. (2021). Synchronous caregiving from birth to adulthood tunes humans' social brain. *Proc. Nat. Acad. Sci. U.S.A.* 118. doi: 10.1073/pnas.2012900118

Zhang, D., Wu, J., Zhang, R., Zhang, H., and He, Z. (2013). High precision micro-impulse measurements for micro-thrusters based on torsional pendulum and sympathetic resonance techniques. *Rev. Sci. Instruments* 84, 125113. doi: 10.1063/1.4850615

Ziada, S., and Lafon, P. (2014). Flow-excited acoustic resonance excitation mechanism, design guidelines, and counter measures. *Appl. Mech. Rev.* 66, 1–22. doi: 10.1115/1.4025788

Zwaan, R. A., and Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *J. Exp. Psychol.* 135, 1–11. doi: 10.1037/0096-3445.135.1.1

**Conflict of Interest:** TM was employed by the company Intheon Labs.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Lomas, Lin, Dikker, Forster, Lupetti, Huisman, Habekost, Beardow, Pandey, Ahmad, Miyapuram, Mullen, Cooper, van der Maden and Cross. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/fnbot.2022.846219

# Avoiding Catastrophe: Active Dendrites Enable Multi-Task Learning in Dynamic Environments

*Abhiram Iyer[1,2†], Karan Grewal[1†], Akash Velu[3†], Lucas Oliveira Souza[1], Jeremy Forest[4] and Subutai Ahmad[1*†]*

[1] Numenta, Redwood City, CA, United States, [2] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, United States, [3] Department of Computer Science, Stanford University, Stanford, CA, United States, [4] Department of Psychology, Cornell University, Ithaca, NY, United States

A key challenge for AI is to build embodied systems that operate in dynamically changing environments. Such systems must adapt to changing task contexts and learn continuously. Although standard deep learning systems achieve state of the art results on static benchmarks, they often struggle in dynamic scenarios. In these settings, error signals from multiple contexts can interfere with one another, ultimately leading to a phenomenon known as catastrophic forgetting. In this article we investigate biologically inspired architectures as solutions to these problems. Specifically, we show that the biophysical properties of dendrites and local inhibitory systems enable networks to dynamically restrict and route information in a context-specific manner. Our key contributions are as follows: first, we propose a novel artificial neural network architecture that incorporates active dendrites and sparse representations into the standard deep learning framework. Next, we study the performance of this architecture on two separate benchmarks requiring task-based adaptation: Meta-World, a multi-task reinforcement learning environment where a robotic agent must learn to solve a variety of manipulation tasks simultaneously; and a continual learning benchmark in which the model's prediction task changes throughout training. Analysis on both benchmarks demonstrates the emergence of overlapping but distinct and sparse subnetworks, allowing the system to fluidly learn multiple tasks with minimal forgetting. Our neural implementation marks the first time a single architecture has achieved competitive results in both multi-task and continual learning settings. Our research sheds light on how biological properties of neurons can inform deep learning systems to address dynamic scenarios that are typically impossible for traditional ANNs to solve.

**Keywords: dendrites, continual learning, reinforcement learning, neuroscience, embodied cognition**

OPEN ACCESS

***Edited by:***
*Adam Safron, Johns Hopkins Medicine, United States*

***Reviewed by:***
*Adeel Razi, Monash University, Australia*
*Ikuko Smith, University of California, Santa Barbara, United States*

***Correspondence:**
*Subutai Ahmad*
*sahmad@numenta.com*

*† These authors have contributed equally to this work*

***Received:*** *30 December 2021*
***Accepted:*** *31 March 2022*
***Published:*** *29 April 2022*

***Citation:***
*Iyer A, Grewal K, Velu A, Souza LO, Forest J and Ahmad S (2022) Avoiding Catastrophe: Active Dendrites Enable Multi-Task Learning in Dynamic Environments. Front. Neurorobot. 16:846219. doi: 10.3389/fnbot.2022.846219*

## 1. INTRODUCTION

Creating embodied systems that thrive in dynamically changing environments is a fundamental challenge for building intelligent systems. Humans handle such environments with ease, but today's deep learning systems struggle with them. Standard Artificial Neural Networks (ANNs) often fail dramatically when learning multiple tasks, a phenomenon known as *catastrophic forgetting* (McCloskey and Cohen, 1989; French, 1999) where the network forgets previously-learned

information. ANNs are inherently designed for static environments with batch training, and learning multiple sequential tasks can lead to significant interference between tasks. Embodied systems, where an agent actively behaves in a changing environment, pose additional challenges. In dynamic scenarios, the training dataset itself is not fixed. Sensory inputs are dependent on an agent's actions and as an embodied agent learns, the actions taken for a given context change as well. Thus, a network learning in these situations needs to avoid forgetting relevant information, update only the information that requires fine tuning, and forget the information that is no longer relevant. The network must distinguish between these types of information categories instead of treating all information as equivalent. The optimal algorithms and architectures for learning in dynamic environments are unknown and remain a fundamental research challenge for AI.

We investigate these questions by looking to neuroscience and biological systems for clues to inform ANNs. In particular we hypothesize that biological properties of *pyramidal neurons* in the neocortex can enable targeted context-specific representations that avoid interference. Most ANNs today rely on an idealized (but inaccurate) model of neurons known as the *point neuron model*, consisting of a linear weighted sum of inputs followed by a non-linearity (**Figure 1**, left). Proposed over a hundred years ago (Lapique, 1907), the point neuron model continues to form the basis for current deep learning systems (McClelland et al., 1986; LeCun et al., 2015). In contrast, pyramidal neurons, which comprise most cells in the neocortex, are significantly more sophisticated and demonstrate a wide range of complex non-linear dendrite-specific integrative properties (Spruston, 2008; **Figure 1**, right). Experimental evidence suggests that dendrites are important for learning task-specific patterns (Yang et al., 2014). In this article we incorporate into an ANN two properties of biological neural networks: active dendrites, and sparsity *via* local inhibition.

We explore the impact of these properties in two non-traditional machine learning scenarios: *multi-task reinforcement learning* (multi-task RL) and *continual learning*. In multi-task RL, a robotic agent learns to perform a diverse set of independent tasks (Yu et al., 2019). Even though tasks are interleaved through training, standard ANNs suffer from significant task interference. In continual learning, a network is trained sequentially on a set of tasks and evaluated on all tasks after training (McCloskey and Cohen, 1989; van de Ven and Tolias, 2019). Here, standard ANNs do not perform well due to catastrophic forgetting. Specifically, because ANNs with point neurons overwrite most of their connections during each iteration of learning, tasks learned at the beginning of training are forgotten and receive low accuracy scores during the evaluation phase (French, 1999; Parisi et al., 2019).

The rest of the article is arranged as follows. After discussing background material, we propose a new architecture that incorporates dendrites and sparse representations into deep learning. We then test our architecture on one representative benchmark from each of the two scenarios, multi-task RL and continual learning. We show experimental results on a standard multi-task RL benchmark, Meta-World. We also show results on a standard continual learning benchmark, permutedMNIST. The results in both cases show that an identical architecture with active dendrites performs well in both benchmarks. Finally, we analyze the results and show that active dendrites and sparse representations help with catastrophic forgetting and gradient interference by learning to create task-specific subnetworks where representations are sparse and mostly orthogonal. Overall, our results suggest that detailed biological properties of neurons can be used to address dynamic scenarios that are difficult for traditional ANNs to solve.

**FIGURE 1 | (Left)** The point neuron prevalent in most ANNs today computes a simple linear weighted sum of its inputs followed by a non-linearity. **(Right)** Morphology of a representative pyramidal neuron. Pyramidal cells in the brain exhibit a vastly more complex structure and functionality. Inset shows a prototypical basal dendritic segment that acts as an independent computational unit.

## 2. BACKGROUND

### 2.1. Multi-Task Learning

The goal of multi-task learning (Caruana, 1997) is to learn a single function that can solve a variety of different learning tasks. The literature in multi-task learning spans many subfields of machine learning, including computer vision (Misra et al., 2016; Kendall et al., 2019; Liu et al., 2019; Purushwalkam et al., 2019), and natural language processing (Dong et al., 2015; McCann et al., 2018). The fields of multi-task RL and continual learning can be seen as subsets of multi-task learning. In the former, tasks are learned in parallel. Conversely, in continual learning, tasks are learned in an ordered sequence.

Compared to single-task machine learning, learning multiple distinct tasks introduces new challenges. When using gradient-based learning algorithms such as backpropagation[1], one challenge is that error gradients and accumulated knowledge from different tasks can interfere with one another. The weight changes necessary to reduce the error for one task may be very different from the changes required for another task. This is a common problem sometimes defined as catastrophic forgetting (French, 1999) or catastrophic interference (McCloskey and Cohen, 1989) in continual learning.

Yu et al. (2020) propose a method to modify conflicting gradients through gradient projection. Several other works demonstrate that using or changing the gradients *via* various normalization, gradient-similarity, and regularization techniques

[1]In this article, the term *backpropagation* refers to the learning method used in deep learning (Rumelhart et al., 1986) and not the phenomenon of back propagating action potentials in dendrites.

can improve learning in multi-task settings (Zhang and Yeung, 2014; Chen et al., 2018; Sener and Koltun, 2018; Du et al., 2020). Novel network architectures are an alternate strategy for avoiding interference in multi-task computer vision settings. Rosenbaum et al. (2018) implement routing networks, learned functions that use task information to determine how to compose a set of function blocks. Liu et al. (2019), Maninis et al. (2019) demonstrate that attention-based architectures could also prevent task interference in multi-task learning scenarios.

#### 2.1.1. Multi-Task Reinforcement Learning

*Reinforcement learning* (RL) is a branch of machine learning in which an agent acts in an environment and receives rewards for each action taken (Sutton and Barto, 2018). The goal is to train an agent, whose actions are determined by a *policy function*, to maximize the total reward received. One fundamental challenge of RL is that the training set itself is highly dynamic. As the agent learns and updates its policy function, it chooses different actions, which in turn changes the sequence of inputs that are received.

*Deep RL* uses deep learning networks to represent the policy function (see Arulkumaran et al., 2017 for a review). Recent years have witnessed the promise of deep RL in a variety of different settings. Mnih et al. (2013) demonstrate that an agent trained with their Deep Q-Network can surpass the performance of expert humans in Atari video games. A few years later, Silver et al. (2018) achieve superhuman performance in more challenging games such as Chess and Go. Other algorithms achieve strong performance in continuous environments with continuous action inputs (Lillicrap et al., 2016). Other methods attempt to induce beneficial learning behaviors such as more stable training (Schulman et al., 2017) and improved exploration (Haarnoja et al., 2018b).

*Multi-task reinforcement learning* combines Deep RL with multi-task learning (Wilson et al., 2007; Yang et al., 2020; Yu et al., 2020). Multi-task RL leads to particularly challenging and interesting scenarios where the system must address both dynamic training regimes and interference from multiple tasks. The idea of separating a neural network into different modules which are composed in a task-dependent manner is proposed in multi-task RL to prevent gradient interference (Andreas et al., 2017; Devin et al., 2017; Sahni et al., 2017; Haarnoja et al., 2018a; Goyal et al., 2020; Yang et al., 2020). Policy distillation, in which information from a "teacher" network is condensed to a smaller "student network," is another popular approach to combine information from different tasks in an effective manner (Rusu et al., 2016).

#### 2.1.2. Continual Learning

While multi-task RL requires the simultaneous acquisition of multiple skills, continual learning requires the sequential acquisition of multiple skills. More generally, continual learning is the ability to acquire new knowledge over time while retaining relevant information from the past. A typical scenario involves training a network on a set of distinct tasks presented in a strict sequence of training phases. Testing the network involves measuring accuracy on all past tasks. van de Ven and Tolias (2019) and Parisi et al. (2019) extensively review the field. Two common approaches to catastrophic forgetting in continual learning involve regularization and subnetworks methods.

Regularization-based methods in continual learning regulate plasticity levels throughout the network during the course of training. In recent years, two of the most prominent examples of regularization are Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and Synaptic Intelligence (SI) (Zenke et al., 2017). Both methods (EWC and SI) estimate the relevance of each weight of the network in solving each task. Inspired by the complex synapse structures seen in biology, SI uses an additional parameter per weight with internal dynamics that depend on the relevance of each weight to each task.

Subnetwork-based methods reduce task interference by identifying subpopulations of neurons that each learn one of the many tasks in the sequence. Gated Linear Networks (Veness et al., 2021) and Dendritic Gated Networks (Sezener et al., 2021) are examples of this type of approach and work by applying a gating mechanism that selects subnetworks based on the input. Context-dependent Gating (XdG) (Masse et al., 2018) selects predetermined subnetworks of neurons, but exact task information must be provided both at training and test times. Similarly, in Wortsman et al. (2020) each task is designated a sparse subset of neurons in the network.

### 2.2. Properties of Biological Neurons

Biological neural networks have evolved in ways that make them much more resilient to catastrophic forgetting and are able to perform significantly better in dynamical scenarios than any ANN to date. ANNs and their component point neurons emerged as simplified abstractions of the complex processes occurring in biological networks and neurons respectively. In this section, we explore the complexities of biological neural networks and review a few properties that are relevant to our work.

#### 2.2.1. Neurons and Active Dendrites

The *pyramidal neuron* is the most prevalent neuron type found in the neocortex and hippocampal areas (Spruston, 2008; Ramaswamy and Markram, 2015). In particular they represent the most common excitatory neuron type found in areas associated with advanced cognitive functions (Spruston, 2008). A typical pyramidal neuron has an extensive dendritic tree containing thousands of synapses, each receiving input from another neuron (y Cajal, 1894; Bentivoglio and Swanson, 2001; Kandel, 2012). The *point neuron model* (Lapique, 1907) postulates that all of these synapses have a linear impact on the cell. This simple assumption formed the basis for Rosenblatt's original Perceptron (Rosenblatt, 1958) and continues to form the basis for current deep learning systems and ANNs (McClelland et al., 1986; LeCun et al., 2015).

Today it is well-known that the point neuron assumption is an oversimplified model of biological computations. Proximal synapses (close to the cell body) have a linear impact on the neuron, but the vast majority of synapses are located on distal dendritic segments (far from the cell body) and individually have minimal impact on the cell. These distal segments process groups of synapses locally in a non-linear fashion, and are known as *active dendrites* (Magee, 2000; Antic et al., 2010; Major et al., 2013;

Stuart and Spruston, 2015; Stuart et al., 2016). Empirical evidence (London and Häusser, 2005; Branco and Häusser, 2010) suggests that each distal dendritic segment acts as a separate active subunit performing its own local computation. Modeling studies show that neurons with active dendrites are more powerful and complex than the point neuron model can accommodate (Poirazi et al., 2003; Jadi et al., 2014; Poirazi and Papoutsi, 2020; Beniaguev et al., 2021).

When input to an active dendritic segment reaches a threshold, the segment initiates a *dendritic spike* (Antic et al., 2010). In basal dendritic segments, dendritic spikes travel to the cell body and can depolarize the neuron for an extended period of time, sometimes as long as half a second (Antic et al., 2010; Major et al., 2013; Gao et al., 2021). During this time, the cell is significantly closer to its firing threshold and any new input is more likely to make the cell fire. This suggests that basal active dendrites have a *modulatory*, long-lasting impact on the cell's response, with a very different role than proximal, or feedforward, inputs (Hawkins and Ahmad, 2016; Antic et al., 2018). Active dendritic segments typically receive contextual input that is a different input than received in proximal segments. These context signals can arrive from other neurons in the same layer, neurons in other layers, or in the form of top-down feedback. Recent experimental evidence has shown that the input on active segments can drive context-dependent activity (Takahashi et al., 2020). In our model, we incorporate these ideas and explore the possibility of using context to create task-specific subnetworks.

#### 2.2.2. Sparse Representations

Neural circuits in the neocortex are highly sparse. Studies show that relatively few neurons spike in response to a sensory stimulus across multiple sensory modalities (Attwell and Laughlin, 2001; Barth and Poulet, 2012; Liang et al., 2019). Sparsity is also present in neural connectivity; cortical pyramidal neurons show sparse connectivity to each other and receive relatively few excitatory inputs from most surrounding neurons (Holmgren et al., 2003). These two phenomena are significantly different from standard ANNs, where both activations and connectivity are dense.

When modeling sparsity in ANNs, sparse neural representations are translated into vectors where most of the entries are off (i.e., equal to zero; Majani et al., 1989). Just like in dense representations, individual entries in a sparse representation can correspond to the presence of certain features (e.g., the unique position of an edge in an input image). One advantage of sparsity in representations is that vectors for two separate entities have low *overlap*, which means the set of features/entries that are non-zero in both vectors is small. Previous studies show that sparse representations are more resistant to noise than dense representations, and slight perturbations in the input are less likely to hinder a trained pattern recognizer (Ahmad and Hawkins, 2016; Ahmad and Scheinkman, 2019; Paiton et al., 2020). The idea of low representation overlap among unrelated inputs may be particularly useful when an ANN is learning multiple, unrelated tasks. If the representations of two different tasks have near-zero overlap, it is possible to significantly reduce task interference. We explore this question in our simulations below.

## 3. ACTIVE DENDRITES NETWORK MODEL

Our primary goal is to augment standard ANNs with the biological properties described above. The extensions should be general and applicable to a range of complex scenarios such as multi-task RL and continual learning. The key aspects of our model are summarized as follows, with details noted in the rest of this section:

1. Pyramidal neurons integrate a range of diverse inputs on multiple independent dendritic segments. To model this, we implement neurons that separate out contextual inputs from feedforward inputs. Each neuron processes the feedforward input using a linear weighted sum. A set of independent dendritic segments process the contextual input using a separate set of weights.
2. Contextual inputs on active dendrites can modulate a neuron's response, making it more likely to fire. To model this, we implement a function that can up-modulate or down-modulate the feedforward activation based on dendritic activation.
3. Neural activity and connectivity are highly sparse. To model this, we incorporate a $k$-Winner-Take-All function ($k$WTA) that mimics biological inhibitory networks (Cui et al., 2017) and guarantees sparse activations.

The above properties are implemented such that the entire network is differentiable and trainable end-to-end using backpropagation. This makes the architecture suitable for testing on any standard deep learning scenario.

### 3.1. Active Dendrites Neuron

Building on the original HTM neuron model (Hawkins and Ahmad, 2016), our *Active Dendrites Neuron* [**Figure 2** (right inset)] receives two sources of input, analogous to the proximal and distal inputs in pyramidal neurons. Feedforward activation is computed by a linear weighted sum of the feedforward input vector, identical to the mechanism in a point neuron. Meanwhile, multiple dendritic segments process a context vector, and the subsequent dendritic output modulates the feedforward activation. This computation produces a neuron where the magnitude of the response to a given stimulus is highly context-dependent.

Given input vector $\boldsymbol{x}$, weights $\boldsymbol{w}$, and bias $b$, our neuron computes the following feedforward activation:

$$\hat{t} = \boldsymbol{w}^\top \boldsymbol{x} + b \tag{1}$$

Similarly, each dendritic segment $j$ computes $\boldsymbol{u}_j^\top \boldsymbol{c}$, given weight $\boldsymbol{u}_j$ and context vector $\boldsymbol{c}$. (The method we use to compute the context vector, $\boldsymbol{c}$, is described in later sections.) We select the segment with the strongest response to the context when computing dendritic activation $d$, which is used to modulate the neuron:

$$d = \max_j \boldsymbol{u}_j^\top \boldsymbol{c} \tag{2}$$

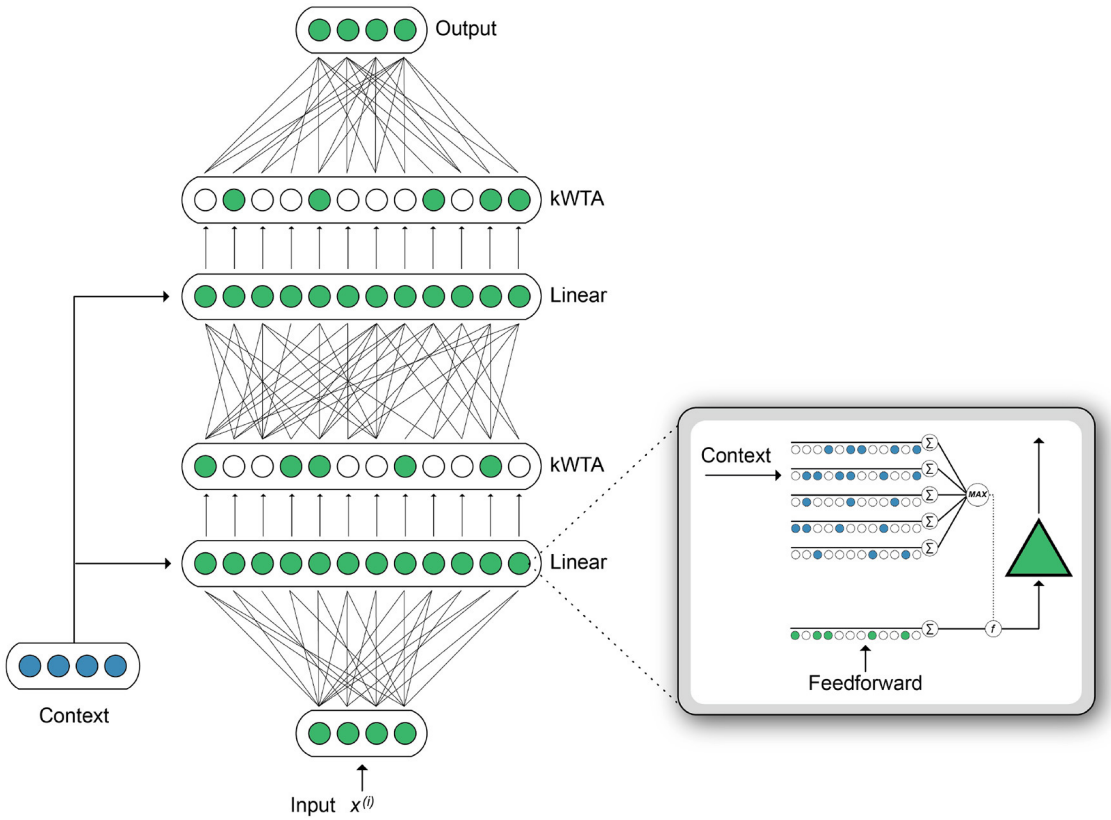



**FIGURE 2 | [Right (inset)]** Illustration of a single Active Dendrites Neuron. Feedforward weights (green) receive regular feedforward input while dendritic segments (blue) receive a context vector. After all dendritic segments compute an activation value, the highest value modifies the linear weighted sum computed by feedforward weights. **(Left)** An overview of the base network structure used in our experiments. There are two hidden layers. Each layer outputs sparse activations, as determined by a *k*WTA activation function. In addition, the weights between layers can be sparse. A context vector is computed for each input. The dendritic segments in each layer receive this context vector as input.

In order to modulate feedforward activation $\hat{t}$ by the dendritic activation $d$, we use modulation function $f(\hat{t}, d)$ where $f(m, n) = m \times \sigma(n)$. Here, $\sigma(\cdot)$ is the sigmoid function which takes a real number and maps it into the range $[0, 1]$. Therefore, by combining (1) and (2) with $f$, we can write the output of a single Active Dendrites Neuron as:

$$\hat{y} = f\left(\hat{t}, d\right) \tag{3}$$

$$= f\left(\boldsymbol{w}^\top \boldsymbol{x} + b, \max_j \boldsymbol{u}_j^\top \boldsymbol{c}\right) \tag{4}$$

$$= \left(\boldsymbol{w}^\top \boldsymbol{x} + b\right) \times \sigma\left(\max_j \boldsymbol{u}_j^\top \boldsymbol{c}\right) \tag{5}$$

Here, a strong positive dendrite response to the context vector will retain the feedforward activation. Conversely, weak or negative responses to the context vector will significantly reduce the activation. We note that there are many variations of (2) that are possible. We found that the network works best when we select the dendrite activation with the largest absolute value and retain the sign in $d$ (Section 6.3).

## 3.2. Sparse Representations

We apply a *k*WTA activation function (Ahmad and Scheinkman, 2019) as our choice of non-linear activation in each hidden layer of the network:

$$k(\hat{y}_i) = \begin{cases} \hat{y}_i & \text{if } \hat{y}_i \text{ is one of the top } k \text{ activations over all } i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $i$ indexes neurons in the same layer. The effect of *k*WTA is to ensure sparsity by selecting the top $k$ activations and setting all others to zero. Feedforward layers that are modulated by dendritic segments and apply *k*WTA thus produce sparse activity patterns that are highly context-dependent. Additionally, our feedforward layers also use sparse weights as proposed in Ahmad and Scheinkman (2019).

## 3.3. Active Dendrites Network Architecture

**Figure 2** (left) shows our Active Dendrites Network, trained end-to-end with backpropagation, where all neurons in each hidden layer are Active Dendrites Neurons. We make two notes: first, only the neurons that were selected by the *k*WTA function will have non-zero activations (and thus non-zero

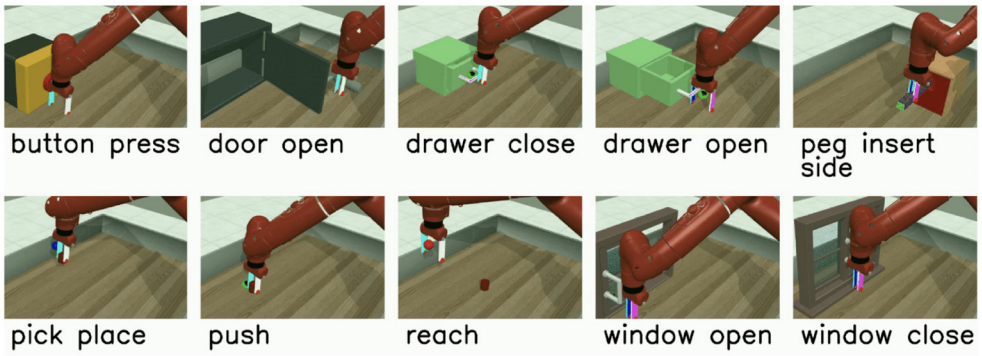



**FIGURE 3 |** The Meta-World v2 Multi-Task 10 (MT10) environment, where a single robotic arm must learn to solve a variety of tasks ranging in difficulty.

gradients). Therefore, during the backward pass, only the weights corresponding to those winning neurons will be updated. Second, for each of those winner neurons, only the dendritic segment $j$ that was chosen by the max operator is updated; all other segments $\boldsymbol{u}_{j'}$ for $j' \neq j$ remain untouched. Thus a very small sparse subset of the full network is actually updated for each input.

We hypothesize that a functional specialization will emerge where different dendritic segments will each learn to identify specific context vectors. Since most dendritic segments that don't respond to a specific context will not be updated, any context-dependent modulation of the neuron should be preserved from task to task. Ideally, the whole process will preserve any context-dependent modulation of a neuron between tasks, reduce gradient interference, and prevent catastrophic forgetting.

# 4. RESULTS

## 4.1. Results With Multi-Task Reinforcement Learning

The multi-task RL problem we investigate uses the Meta-World v2 environment and its associated v2 tasks (Yu et al., 2019). Meta-World contains multiple different object manipulation tasks that a single robotic arm must learn to solve simultaneously. We use the MT10 environment, which contains 10 tasks ranging in complexity as depicted in **Figure 3**. Although the concrete outcome of each task is unique, all tasks share a common structure that enables the agent to leverage some shared information during training. For instance, learning how to grasp an object is a shared concept among many of the tasks.

The algorithm we use to train our robotic agent is multi-task Soft Actor-Critic (MTSAC) as introduced by Yu et al. (2019), an adaptation of the popular Soft Actor-Critic (SAC) framework proposed earlier by Haarnoja et al. (2018b). MTSAC is an actor-critic deep RL algorithm that maximizes an agent's cumulative reward to solve a task while also maximizing entropy to encourage environment exploration. To maintain consistency with the codebase of Yu et al. (2019) which fixes goal states (e.g., position of an object in the environment) through training, we also keep goal locations constant across all our experiments. A deeper explanation about our multi-task RL setup and the algorithm we use to train the agent can be found in Section 6.1. As in many RL problems, there is no static training and testing dataset. Rather, past experiences from the agent are used to iteratively train the agent. We freeze the network at regular intervals to test accuracy on all tasks.

### 4.1.1. Network Structure for Multi-Task RL

**Figure 4** shows our network architecture for multi-task RL. We use a network with 2 hidden layers—each with 2,800 neurons and followed by a $k$WTA activation function—and a final output layer. The first hidden layer has standard neurons whereas the second hidden layer contains Active Dendrites Neurons which are modulated by the context vector. The primary feedforward input to the network is a state vector consisting of the agent's position in the world as well as the position and orientation of the target object. The output of the network is an action vector that describes the joint torques and gripper forces of the robotic arm. The structure of the state and output vectors is identical across all tasks. All feedforward weights are sparse.

For our multi-task RL experiments, the context vector $\boldsymbol{c}$ encodes the task ID as a one-hot encoded vector. We considered other options to generate $\boldsymbol{c}$, such as first pre-processing the one-hot encoding by a linear layer, but found that a one-hot encoding was adequate. Each Active Dendrites Neuron in our network has exactly 10 dendritic segments (same as the number of tasks to learn) so that each segment can potentially learn to recognize a unique context vector.

We compare our Active Dendrites Network to baselines reconstructed[2] from Yu et al. (2019) which are multi-layer perceptrons (MLPs) with dense weights and ReLU activations. These MLP baselines are used to model both the policy and the Q function. Additionally, these MLP baselines receive context information $\boldsymbol{c}$ in the form of feedforward input concatenated to the state vector. Thus, both the Active Dendrites Network and the baseline network receive *identical* information at each time step;

[2]We are unable to directly present the published baseline results because their plots contain inconsistencies between success rates per-task and across all tasks. To present a fair comparison, we re-run the baseline networks using their codebase and hyperparameters.

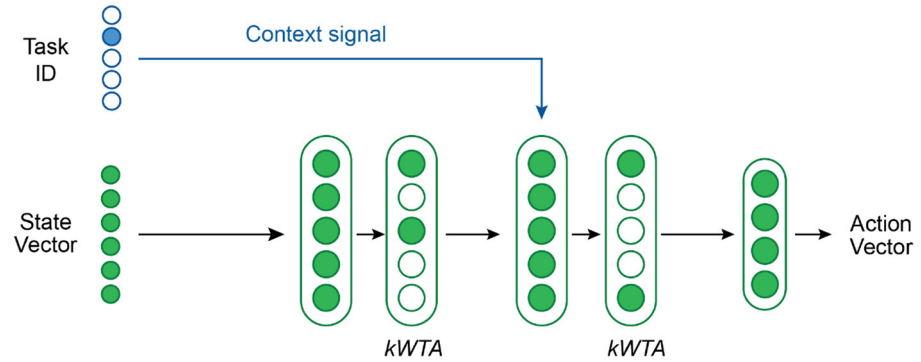



**FIGURE 4 |** An overview of the network structure used in our multi-task RL experiments. A *k*WTA activation is applied to both hidden layers. The context vector is the task ID. The dendritic segments in the second hidden layer receive this context vector as input.

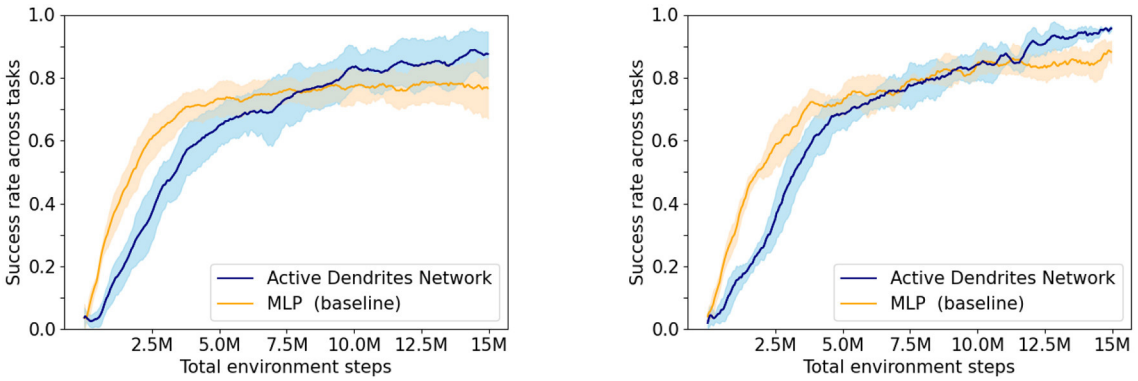



**FIGURE 5 |** The success rate of our network when learning 10 tasks compared to the MLP baseline with context. **(Left)** Experiment 1—the average of 10 Active Dendrites Network runs and 10 MLP baseline network runs that all share the same training hyperparameters. **(Right)** Experiment 2—the average of the five best Active Dendrites Network experiments and the five best MLP baseline experiments. The shaded region in each plot represents the standard deviation of the success rate from the average.

the primary difference between the two architectures is how the context vector is handled.

**Table 2** in Section 6.1.4 shows the networks we ran, the number of *non-zero parameters* in each network, and the hyperparameters used to train each network. Although we control the hidden sizes to yield approximately the same number of total non-zero parameters across our experiments, we note that the MLP baseline network contains nearly 500,000 more non-zero parameters than our Active Dendrites Network. We chose a network with two hidden layers to draw fair comparisons with the MLP baselines presented in Yu et al. (2019). **Supplementary Materials** (Section 1) includes the results of additional experiments detailing the impact of some of our architectural choices.

### 4.1.2. Dendrites Improve Multi-Task RL Accuracy

We show results from two different experiments that compare our Active Dendrites Network to the MLP baseline network.

*Experiment 1*: In this experiment, we assess the overall performance of each architecture. We ran both an Active Dendrites Network and a MLP baseline network with identical training hyperparameters. **Figure 5** (left) shows the mean overall success rate for each architecture during the course of training over 10 independent trials. To identify which architecture performs the best for each task, we compute the mean success rate per task for the last 500,000 environment steps of training and list these values in **Table 1**. Additionally, we show the per-task training statistics during this same segment of training, as seen in **Figure 6**.

We see in **Figure 5** (left) that although the Active Dendrites Network has lower success rates early in training, it overtakes the baseline architecture and is about 10% better by the end. **Table 1** shows that the average end success rate for the Active Dendrites Network (across the last 500,000 steps of training) is **87.5%**. In comparison, the average success rate for the MLP baseline is **76.6%**. We also note that the *push*, *peg-insert-side*, and *pick-place* tasks were the hardest to solve because they are the most unlike the other tasks. Specifically, these three tasks require that a robot grasp and move a small object to a specified location. As evident in **Figure 6** for these three tasks, the median success rate of an

**TABLE 1 |** The mean per-task success rate produced by each network in Experiment 1.

| Tasks | Model | |
|---|---|---|
| | Active Dendrites Network (%) | MLP baseline (%) |
| Drawer-close | **100.0** | **100.0** |
| Window-close | **99.7** | 95.3 |
| Button-press-topdown | 95.7 | **97.3** |
| Reach | **99.7** | 86.3 |
| Window-open | **99.3** | 93.7 |
| Drawer-open | **99.7** | 86.0 |
| Door-open | **94.7** | 84.3 |
| Push | **67.3** | 59.0 |
| Peg-insert-side | **71.7** | 47.0 |
| Pick-place | **47.7** | 17.3 |
| Overall | **87.5** | 76.6 |

*The success rates are averaged over the last 500,000 steps of training. The best success rate between the Active Dendrites Network and MLP baseline is highlighted in bold.*

Active Dendrites Network is far greater than that of the MLP baseline network. We hypothesize that these tasks are hard to learn because of significant gradient interference with the other tasks, and that the context-specific sparsity imposed by the Active Dendrites Network helps remove this interference.

*Experiment 2*: The high variance in **Figure 5** (left) is inherent in many RL scenarios (Irpan, 2018; Ibarz et al., 2021). This is in large part due to the highly stochastic and dynamic training process. For instance, small variations in the trained policy can result in large variations in the agent's behavior which significantly impacts the data collected during training. Additionally, a policy can generate different behaviors during training when sampling from its predicted action distribution.

To control for some of this variation in training, for each network initialization we select the best result across different training runs. For each of five different Active Dendrites Network initializations, we ran five training runs and picked the run with the highest end success rate across the last 500,000 environment steps of training. We then compute the mean overall success rate across these five best runs. We follow the same procedure for finding the five best MLP baseline networks and compare the results in **Figure 5** (right).

We find that this process significantly reduces the variance and that the best Active Dendrites Networks still outperform the best MLP baseline networks. Across the five best Active Dendrites Network runs, the average overall end success rate is **95.6%**. In comparison, across the five best MLP baseline runs, the average overall end success rate is **88.2%**.

## 4.2. Results With Continual Learning

A typical continual learning problem consists of training a neural network on a discrete number of tasks in sequence. Once a network is trained on a particular task, it does not encounter that task during training again. The goal is to learn all the tasks in sequence without forgetting previously-learned tasks.

We use the permutedMNIST dataset (Goodfellow et al., 2014), a common benchmark in continual learning where each task requires classifying images of handwritten digits (0–9) after a unique pixel-wise permutation has been applied. Since the data distribution of each task changes and because neural networks are generally not permutation-invariant, forgetting occurs.

We use the original MNIST training dataset of 60,000 images to construct the dataset for a single task. Since we train on $\mathcal{T}$ consecutive tasks, the network is trained on a total of $\mathcal{T} \times 60,000$ images. Once training is complete, the network accuracy is calculated using a test set consisting of all $\mathcal{T}$ permutations applied to the MNIST test dataset of 10,000 images.

We train our model to learn up to 100 tasks in sequence. The network is tested at the end of training by computing accuracy on the test set for all tasks. When attempting to learn $\mathcal{T}$ consecutive tasks, the hidden neurons are equipped with $\mathcal{T}$ dendritic segments each to give it sufficient capacity to recognize a unique context vector for each task. We report accuracy numbers by averaging over 8 independent runs each with a randomly-picked seed. See Section 6.2 for additional details.

### 4.2.1. Computing the Context Vector

As with multi-task RL, we need to compute an appropriate context vector. For continual learning, we use a simple prototype method (Rosch, 1975; Snell et al., 2017) to select the context vector where a single vector represents each task [**Figure 7** (left)]. We implement two different variations of the prototype method depending on the knowledge available to the system during training.

#### *4.2.1.1. Training Method 1 (Task Information Provided)*

In the first method, we assume that the system receives task information during training, when all training samples for a particular task are assigned a single prototype context vector. We compute the prototype vector for task $\tau$ by taking the element-wise mean over all the training samples across all features:

$$\boldsymbol{p}_\tau = \frac{1}{|V_\tau|} \sum_{\boldsymbol{x} \in V_\tau} \boldsymbol{x}$$

where $V_\tau$ denotes the set of all data samples $\boldsymbol{x}$ that the model observes to train on task $\tau$. The dimensionality of the context vector is thus identical to the dimensionality of the input vectors. This context vector is specific to each task and agnostic to the target label.

#### *4.2.1.2. Training Method 2 (Task Information Not Provided)*

In the second method, we relax the constraint that the identity of the task is given during training and instead implement prototypes that are automatically selected during training. To achieve this, we use a statistical clustering approach that builds context prototypes on the fly. When the system receives a new batch of training samples from a task, we use an unpaired multivariate $t$-test to compare the current samples to previously-observed training samples. If the new batch of samples is similar to earlier training samples, they are assigned to an existing prototype. If not, the new batch of samples is assumed to

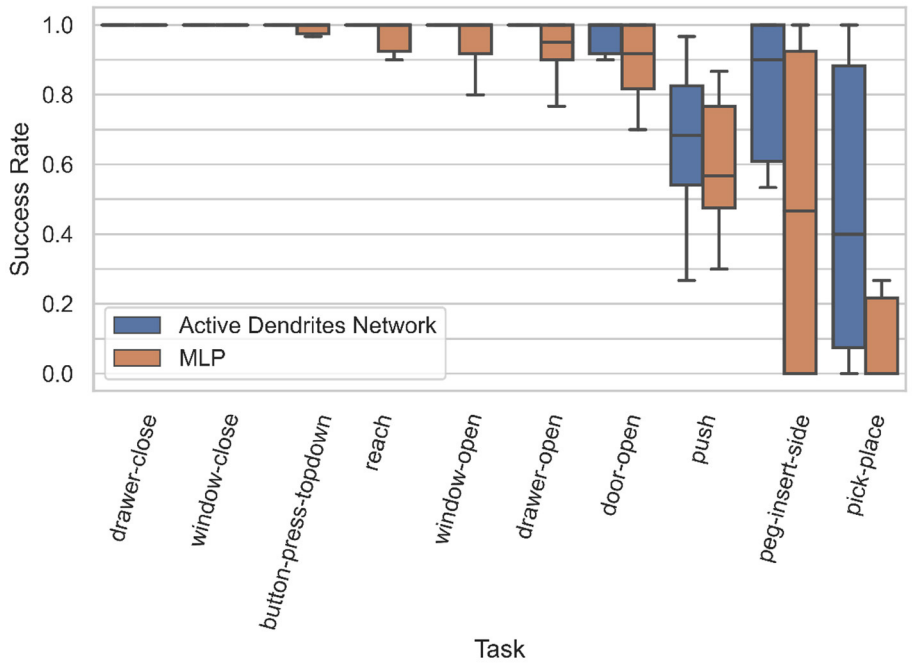



**FIGURE 6 |** Box plots of the accuracies for each MT10 task for our Active Dendrites Networks and MLP baseline networks in Experiment 1. We discard outliers for all runs for clarity.

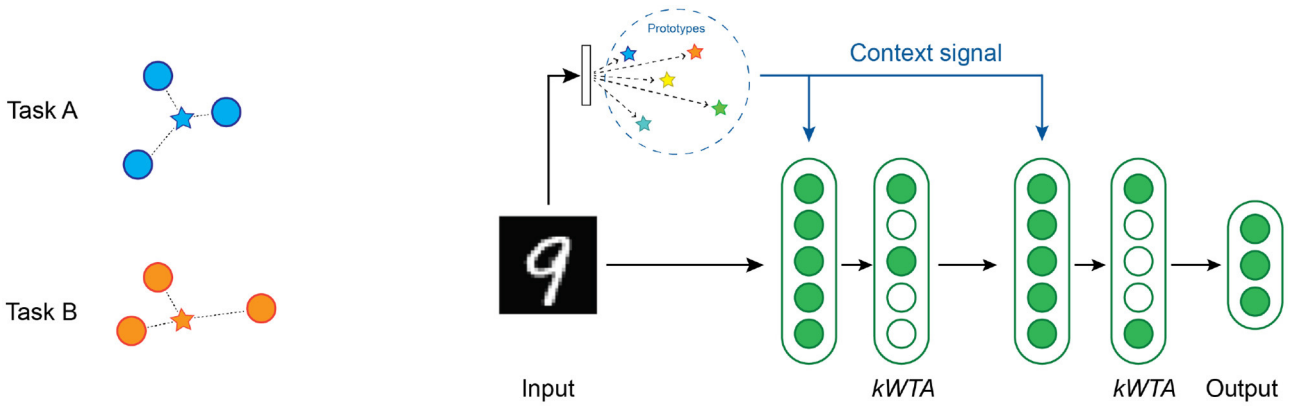



**FIGURE 7 | (Left)** An illustration of the prototype method for computing the context vectors. The blue circles are training samples in input space for task A, while the orange circles are training samples for task B. The blue star is a vector that represents the prototype for task A, and the orange star represents the prototype for task B. **(Right)** An overview of the network structure used in our continual learning experiments. There are two layers of hidden units, each with a *k*WTA activation function. A context vector is computed from each image by locating the nearest prototype vector.

correspond to a new task, and a novel prototype is instantiated. In this case, there isn't necessarily a one-to-one mapping between tasks and prototype context vectors. More details on this method are described in Section 6.2.3.

#### *4.2.1.3. Selecting Prototypes During Inference*

For both methods above, we do not provide any task information to the system during evaluation. Instead it must dynamically select the correct context vector and provide that to the network. We enable this dynamic approach by selecting the closest prototype vector to each test example using Euclidean distance. That is, for a test example $\boldsymbol{x}'$, the chosen prototype is:

$$\arg\min_{\boldsymbol{p}_\tau} ||\boldsymbol{x}' - \boldsymbol{p}_\tau||_2$$

computed over all prototypes $\boldsymbol{p}_\tau$ stored in memory.

### 4.2.2. Network Structure for Continual Learning

**Figure 7** (right) shows the network that we used for our continual learning experiments. Each of the two hidden layers contain

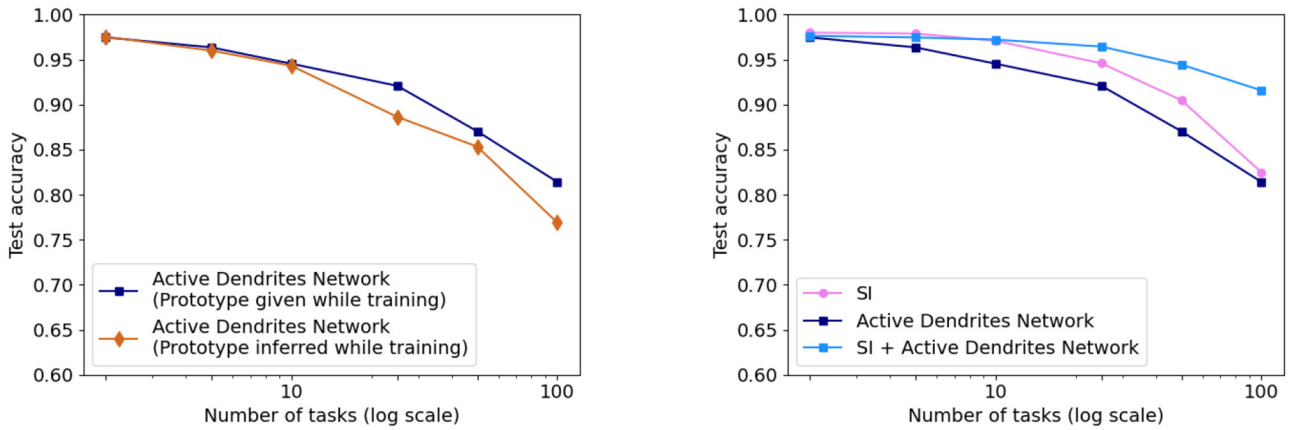



**FIGURE 8 | (Left)** The accuracy of our Active Dendrites Networks when learning 2, 5, 10, 25, 50, and 100 permutedMNIST tasks in sequence. We show results using both prototype methods while training: when the model is provided with a prototype, and when it must select the vector in an online manner. **(Right)** The accuracy of the Active Dendrites Network and SI. The accuracy when combining SI + active dendrites is greater than either one on its own.

2,048 Active Dendrites Neurons followed by a *k*WTA activation function. The output of the network is a standard layer with 10 neurons. We choose our network layer sizes to be similar to previous studies that report results on this dataset (Kirkpatrick et al., 2017; Zenke et al., 2017; Masse et al., 2018). (Section 6.2 details the hyperparameters used for each experiment).

#### 4.2.3. Dendrites Mitigate Catastrophic Forgetting in Continual Learning

As shown in **Figure 8** (left), we achieve accuracies of **94.6** and **81.4%** on 10 and 100 consecutive permutedMNIST tasks, respectively, when context is provided during training, and accuracies of **94.3** and **76.9%** when context is dynamically chosen during training. Since there are always 10 categories, chance accuracy is 10% independent of the number of tasks. This demonstrates that the network successfully retains the majority of the knowledge from previous tasks. Note that a standard feedforward network performs poorly on this benchmark (Kirkpatrick et al., 2017; Zenke et al., 2017; van de Ven and Tolias, 2019; see also Section 4.3.3 for more direct comparisons).

We also compare the results with SI (Zenke et al., 2017; see Section 2.1.2). SI is inspired by the complex structure of biological synapses and known to do well on this benchmark. SI operates solely at the level of synapses: it maintains an additional parameter per weight that controls the speed of weights adapting to specific tasks. In SI, the weight updates are sprinkled throughout the network and not grouped according to units or dendrites. On the other hand, the dendrites in our network impact a small subset of the neurons, and only the weights on these neurons and dendrites are modified. As such, our two approaches seem to be complementary. **Figure 8** (right) shows the benefits of combining these two techniques. The accuracy of Active Dendrites Networks combined with SI improves to **97.2** and **91.6%** accuracy on 10 and 100 consecutive tasks, respectively. Combining the two leads to higher accuracy than either method on its own. This suggests that biological mechanisms at the synapse, neuron, and network levels can operate together to handle continual learning. Note that SI as described in Zenke et al. (2017) requires knowledge of the task during training; therefore we only combine it with our first prototype method. It may be possible to remove this restriction, which is a direction for future research.

#### 4.2.4. Comparison With Context Dependent Gating

The idea of leveraging sparse representations and subnetworks within an ANN to combat catastrophic forgetting is not entirely novel. The implementation closest to ours is XdG (Masse et al., 2018) that uses a hard-coded distinct subnetwork for each task. When training on a task, the implementation invokes the task-specific subset of the hidden layer of the ANN; other neurons are forced to have an activation value of zero. The XdG implementation requires a task ID that determines exactly which neurons to turn on or off. Training Active Dendrites Networks in a continual learning scenario also yields subnetworks and sparse representations. However, we emphasize two major distinctions between our model and XdG:

1. Task information is inferred in our system (*via* prototyping) whereas XdG provides the system with a task ID during training and testing. As such, our system is solving a problem that is known to be significantly more challenging (van de Ven and Tolias, 2019).
2. Subnetworks automatically emerge *via* the use of dendritic segments for each new task whereas XdG pre-allocates a different subnetwork for each task, which also indicates our system is solving a more challenging problem.

We compare Active Dendrites Networks to XdG in **Figure 9**. Just as we augment Active Dendrites Networks with SI, so too does XdG. Our results with a large number of tasks are significantly better than XdG, and slightly worse than XdG combined with SI, but without their limitations.

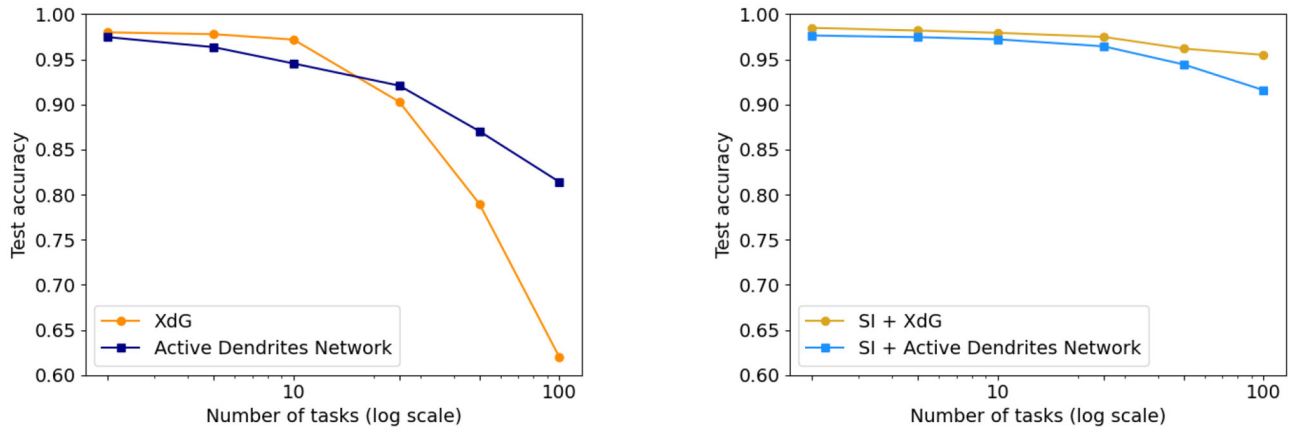



**FIGURE 9 | (Left)** Final accuracy of the Active Dendrites Network in comparison to XdG when learning 2, 5, 10, 25, 50, and 100 permutedMNIST tasks. The more tasks learned by the system, the greater the accuracy of the Active Dendrites Network. **(Right)** Final accuracy of each method when augmented with SI, and SI itself. XdG results are taken from Masse et al. (2018).

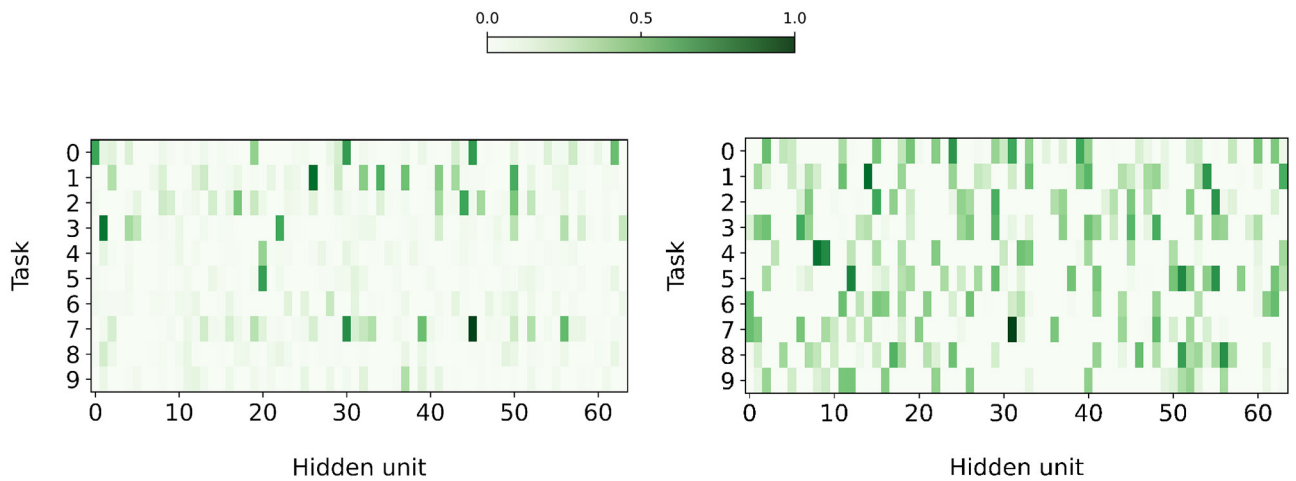



**FIGURE 10 |** The fraction of instances for which each of the first 64 hidden units in the hidden layer became active (after applying *k*WTA), when training an Active Dendrites Network on MT10 tasks **(Left)** and 10 permutedMNIST continual learning tasks **(Right)**. Both figures separate instances by task. For MT10, the figure tests the trained RL policy on each task three times during evaluation. For permutedMNIST, the figure uses 5,000 randomly-chosen test examples across all tasks. Note that each hidden layer contains more than 2,000 hidden units, but we show just 64 for ease of visualization.

Learning is more challenging in our system as dendritic segments must learn the mapping between context vectors and different subnetworks. In effect, sparse representations and minimally overlapping subnetworks emerge organically in our model. We note that perhaps this makes learning more effective as dendritic segments can choose subnetworks that overlap more for tasks that are more semantically related, thus requiring less network capacity.

## 4.3. Analysis

### 4.3.1. Are Dendrites Invoking Subnetworks?

The hypotheses of our work are two-fold. First, Active Dendrites Networks modulate an *individual* neuron's activations for each task. Second, *k*WTA activations use this modulation to activate subnetworks that correspond to each task. To test these hypotheses, we train and analyze an Active Dendrites Network for 10 tasks in multi-task RL and continual learning scenarios and investigate the representations of a layer of neurons modulated by dendritic segments.

**Figure 10** shows the average activation frequency per task (after applying *k*WTA) for the first 64 neurons in the second hidden layer for both multi-task RL and continual learning. Looking horizontally across the rows, each task appears to select a different sparse subset of neurons. Looking vertically across the columns, each neuron appears to activate frequently only for a small fraction of tasks. According to this measure, it appears that the network has indeed learned to invoke minimally overlapping subnetworks for different tasks.

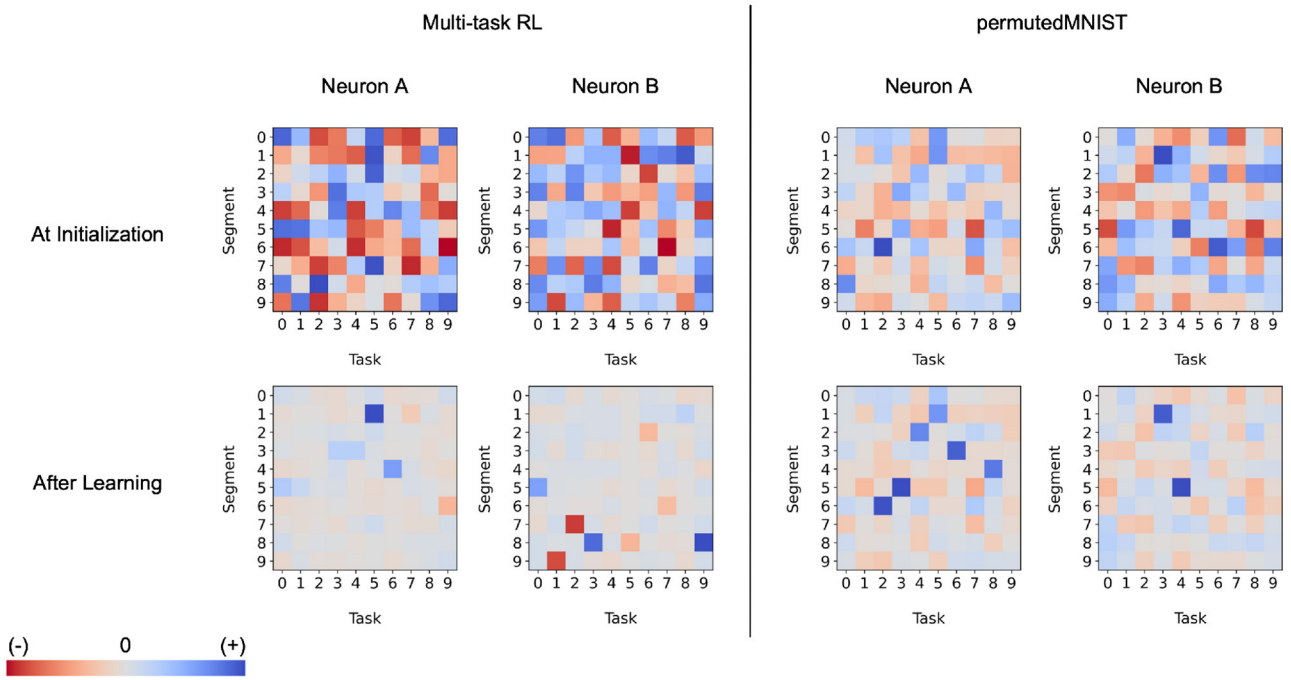



**FIGURE 11 |** The behavior of the dendritic segments of two separate neurons in a hidden layer of an Active Dendrites Network during three random evaluations of each MT10 task and 5,000 random evaluations of each permutedMNIST task. These charts show the activation computed by each dendritic segment given the context vector corresponding to each task, before **(Top)** and after **(Bottom)** training. Note that the dendritic segments for a particular neuron are completely separate of the segments of another in both multi-task RL and continual learning scenarios (e.g., Neuron A's first segment is unrelated to Neuron B's first segment).

What is the effect of dendrites on a single neuron? In **Figure 11**, we analyze a few Active Dendrites Neurons and their responses to different context vectors before and after learning 10 multi-task RL and permutedMNIST tasks in sequence. At the beginning of training, the responses are random with scattered positive, negative, and near-zero responses. After training, most responses are weak and only a few are either strongly positive or negative. Notably, across the neurons, dendrites only have strong responses to a few contexts as different neurons participate in different subnetworks. We note that in the multi-task RL scenario, we observe both strong positive and negative responses while the continual learning scenario only shows strong positive activity. We are unclear as to why this particular behavior emerges in continual learning but not multi-task RL.

### 4.3.2. Impact of Sparsity Level and the Number of Dendrites

We show that an Active Dendrites Network is competitive with benchmarks in both multi-task RL and continual learning. However, to what extent are active dendrites and sparse representations both contributing factors toward alleviating catastrophic forgetting?

We investigate this question in the context of continual learning. We find that both active dendrites without sparse representations and standard point neurons with sparse representations are better than chance in a continual learning scenario. However, the combination of both active dendrites and sparse representations yield significantly better results than either one on its own. As **Figure 12** (left) shows, the accuracy of both methods evaluated independently and evaluated together on 10 and 100 permutedMNIST tasks demonstrates the importance of implementing both active dendrites and sparse representations.

To further test the impact of dendrites and sparsity, we run two additional tests in the continual learning scenario. First, we fix the level of sparsity in our hidden representations and vary the number of dendritic segments per hidden neuron. Second, we fix the number of dendritic segments per hidden neuron and vary the sparsity in our hidden representations (i.e., vary $k$ in $k$WTA). As seen in **Figure 13** (left), increasing the number of dendritic segments leads to a small monotonic increase in accuracy. **Figure 13** (right) shows that reducing sparsity translates to a sharp drop in accuracy, further highlighting the need for sparse representations.

### 4.3.3. Are Networks With Dendrites Equivalent to Larger Networks?

Over the last couple of decades, multiple studies have suggested that dendritic computations performed by pyramidal neurons can be approximated by ANNs that have one or more hidden layers. For example, Poirazi et al. (2003) shows that a larger two-layer neural network can well-approximate the post-synaptic responses of a pyramidal neuron with active dendrites. Various follow-up studies also make similar claims (Jadi et al., 2014; Beniaguev et al., 2021), with Beniaguev et al. (2021) suggesting

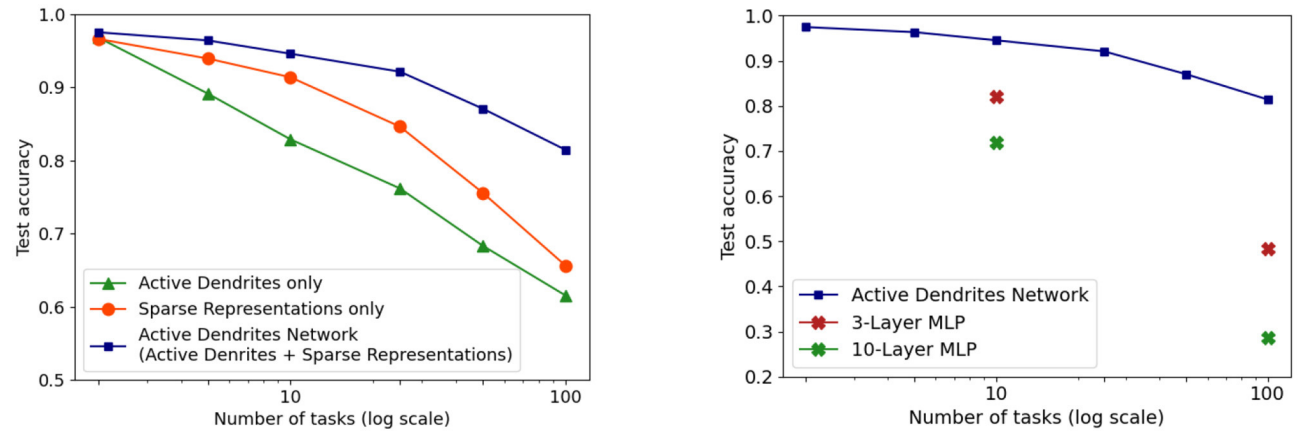



**FIGURE 12 | (Left)** Continual learning test accuracy on permutedMNIST using active dendrites and dense representations (green), regular ANNs with sparse representations (orange), and Active Dendrites Networks (blue) which use both active dendrites and sparse representations. **(Right)** Continual learning test accuracy for our Active Dendrites Network compared to regular feedforward networks with more layers. Our Active Dendrites Network has three layers; the two hidden layers contain neurons modulated by dendritic segments. In all experiments (left and right subfigures), we average results over 8 independent runs, each with a randomly initialized seed, and omit standard error bars as they highlight a very small range.

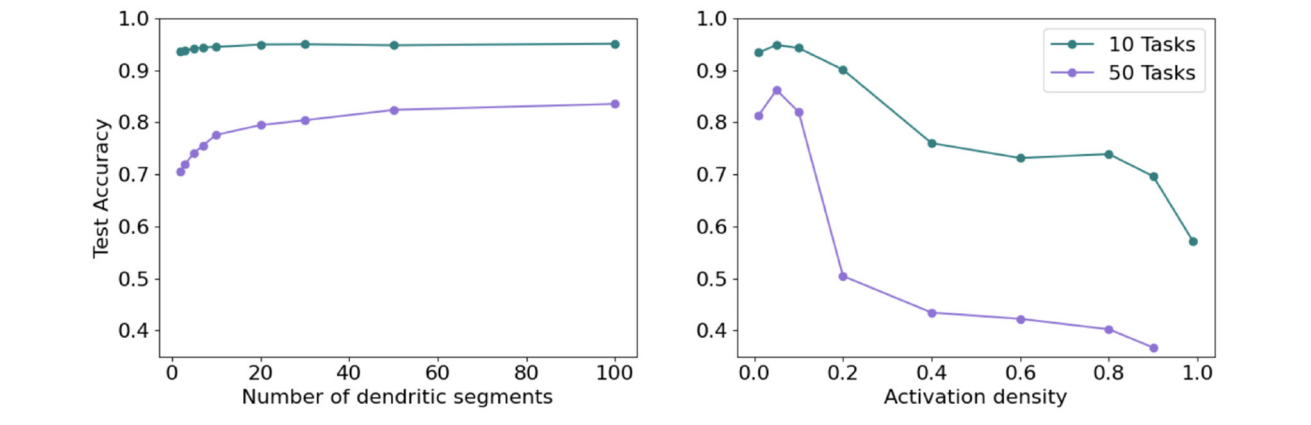



**FIGURE 13 | (Left)** Final accuracy on test examples across all tasks when varying the number of dendritic segments per neuron and keeping activation sparsity constant when learning 10 (top) and 50 (bottom) permutedMNIST tasks. **(Right)** Final accuracy on test examples across all tasks for a fixed number of dendritic segments per neuron and varying activation density level on 10 (top) and 50 (bottom) permutedMNIST tasks.

that a pyramidal neuron is equivalent to a larger ANN with seven hidden layers. In this section we show that in the dynamic scenarios considered here, an Active Dendrites Network is not equivalent to larger or deeper ANNs.

In the case of multi-task RL, a pyramidal neuron's activity *cannot* be approximated by a neural network with more parameters. For instance, classical deep networks that are trained on a variety of tasks are incapable of performing well due to gradient interference, an issue that cannot be solved with simply more hidden neurons. When comparing a three-layer Active Dendrites Network and a three-layer MLP with 500,000 more learnable, non-zero parameters, **Figure 5** shows that networks with dendrites and sparse representations far outperform the MLP baseline. We also experiment with larger 3-layer MLPs that have 1,700,000 more non-zero parameters than our Active Dendrites Network (hyperparameters found in **Table 2** of Section 6.1.4). In this case, we find that the MLP produces a success rate of **73.1%** across 10 tasks (averaged over the last 500,000 environment steps of training), which underperforms our Active Dendrites Network yielding an average success rate of **87.5%**.

In addition, in the continual learning setting, our network with dendrites cannot be approximated by a neural network with multiple layers. When considering continual learning, classical deep networks are incapable of performing well due to catastrophic forgetting, regardless of network depth. This specific trend can be observed in **Figure 12** where our Active Dendrites Network outperforms standard MLPs that have (a) the same number of layers but no dendrites (for 10 and 100 permutedMNIST tasks), and (b) many

more layers and roughly the same number of learnable parameters (for 10 permutedMNIST tasks). (Other ablation studies, not shown in **Figure 12**, are described in the **Supplementary Materials** section.)

These results for both multi-task RL and continual learning suggest that standard ANNs that are wider or deeper are still prone to gradient interference and catastrophic forgetting while active dendrites can help retain knowledge from previous tasks. In these dynamic settings, our experiments show that a standard feedforward network with more hidden units or additional layers is not as powerful as a network with active dendrites.

## 5. DISCUSSION

The exact mechanistic details of how a biological neuron converts incoming signals into action potentials (i.e., spikes) remain unclear. Ever since Rosenblatt (1958), models of biological neurons favor a single linear weighted sum (the point neuron) as a tractable abstraction. This idea continues to serve as the prevalent paradigm in machine learning today for the individual computational unit. One shortcoming is that standard ANNs with point neurons can suffer from catastrophic forgetting. They overwrite many of their connections for each learning iteration, and thus quickly lose previously-acquired knowledge (French, 1999; Parisi et al., 2019).

In this article we show that augmenting point neurons with biological properties such as active dendrites and sparse representations significantly improves a network's ability to learn multiple tasks at once. In the multi-task RL setting, a three-layer Active Dendrites Network can achieve an average accuracy of about **88%** when learning 10 Meta-World tasks together. In the continual learning setting, an almost identical network can achieve greater than **90%** accuracy when learning 100 permutedMNIST tasks in sequence. These results, on two very different scenarios, suggest that Active Dendrites Networks may represent a general purpose architecture for avoiding interference and forgetting in complex settings. In the rest of this Discussion we elaborate on this idea and describe some relationships to other research.

### 5.1. Dendrites Enable Dynamic Context Integration and Routing

In this section, we attempt to elucidate how active dendrites help in dynamic scenarios such as multi-task and continual learning, and discuss our theory of their underlying role in the neocortex. Following experimental evidence (Section 2.2.1), our model suggests that dendritic segments in each neuron identify specific contexts and then modulate neuronal activity based on this identification. Combined with subsequent local inhibition (*k*WTA function), the modulation can impact whether the neuron activates.

We propose that the consequence of this behavior is to invoke sparse context-specific subsets of the network. Two different context vectors can lead to different winners and different sparse activation patterns (illustrated in **Figure 14**). As suggested by the figure, the same feedforward input can activate completely different neurons based on the specific context. Note that the subnetworks are distributed and that two different subnetworks may share some neurons. In Section 4.3.1, we showed that task-specific representations do indeed emerge in our experiments (**Figure 10**).

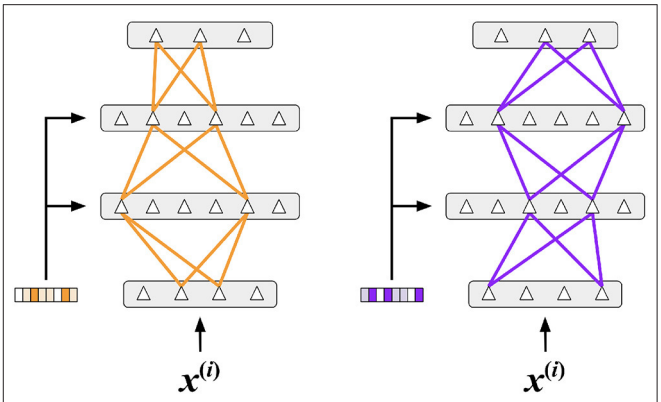



**FIGURE 14 |** A representation of subnetworks within an Active Dendrites Network. By receiving different context vectors as input, dendritic segments invoke different subnetworks for a fixed feedforward input. The subnetworks are distributed, i.e., they may share some of the same neurons.

Why do subnetworks help? In dynamic conditions, the system must react and learn in constantly changing situations. Subnetworks restrict the flow of information to be highly context-dependent and relevant to each specific situation. In addition, errors will only propagate through the active subnetwork. Only the active neurons will update their feedforward weights and only the winning segment within those active neurons will update their dendritic weights. Thus, by utilizing context the brain can isolate information flow, and direct learning itself in a highly localized and task specific manner. The last two decades have seen significant experimental support for highly localized task specific learning in the dendrites of pyramidal neurons (Losonczy et al., 2008; Yang et al., 2014; Kerlin et al., 2019; Limbacher and Legenstein, 2020).

What is the role of context? In this article we have used a context vector that represents the current task. Prior experimental and modeling work shows the utility of various other types of context. In recurrent networks, it is possible to use the previous activity of the network as context for dendrites. In this case a layer of neurons becomes a powerful sequence memory system (Hawkins and Ahmad, 2016). For sensorimotor inference, if the coordinates of an external reference frame is used as context, neurons can perform object recognition with actively moving sensors and by integrating information over time (Hawkins et al., 2017). Schmidt-Hieber et al. (2017) and Heald et al. (2021) also provide experimental evidence for the role of dendrites in separating out information in continuous sensorimotor streams. In the neocortex, if the inference results of neighboring cortical areas are used as context, dendrites can be used to disambiguate uncertain information and perform voting (Hawkins et al., 2017).

In each of the above scenarios, although the nature of the context greatly impacts emerging behavior, the fundamental operations remain the same. Dendrites recognize patterns that best match their synapses, and up-modulate their neurons such that they are more likely to win. This in turn invokes context specific subnetworks that route information flow and gate learning in order to effectively learn and perform the task at hand. Contextual routing mediated by dendrites may thus be a general-purpose and powerful capability that underlies much of cognitive function (Phillips, 2015; Phillips et al., 2015). Indeed, the ability to generate context-dependent output based on a common set of operations could be a crucial building block of cognitive maps able to cover any domain (Whittington et al., 2022). Flesch et al. (2022) provide experimental evidence for contextual gating in a study of human continual learning and memory.

In our implementation we have focused primarily on feed-forward information flow and basal dendrites, and have ignored recurrent and feedback connections and apical dendrites (Larkum et al., 1999; Larkum, 2022). Interestingly, lateral connections and feedback connections seem to segregate onto different dendritic integration zones (Guest et al., 2021; Lafourcade et al., 2022). Recent experimental evidence suggests that apical dendrites also process feedback context and have a modulatory impact on the cell leading to task specific functionality (Kerlin et al., 2019; Takahashi et al., 2020; Schoenfeld et al., 2022). From a modeling perspective, there is additional complexity related to generating top-down context (Siegel et al., 2000) and simultaneously processing three separate input streams (Phillips, 2015; Larkum, 2022), an interesting area for future research.

In this article, we have focused on modeling the dendritic properties of pyramidal neurons, but we note that dendritic modulation and gating may occur with other neuron types. For example, thalamocortical neurons may exhibit analogous dendrite initiated gating properties (Errington and Connelly, 2011). As such, dendrite mediated contextual integration and gating may be a more general phenomenon of biological neural systems. Modeling other neuron types is an interesting area for future work.

## 5.2. Comparison to Other Multi-Task RL Systems

Many techniques in multi-task RL make manual changes to the network structure or learning scheme in order to account for the learning of new tasks. In multi-task scenarios, optimizers struggle to learn different tasks that vary in gradient magnitude and have conflicting gradient direction. In these cases, tasks with larger magnitudes are usually preferred during optimization over others. To rectify this issue, Yu et al. (2020) minimizes gradient interference by orthogonally projecting the gradients of tasks that conflict with each other. Additionally, in most scenarios, a policy trained on a specific task with a specific agent cannot be adapted to similar problem settings. Devin et al. (2017) proposes a framework to learn separate policy modules corresponding to a particular task or robotic agent. Ultimately, they show how these modules can be mixed to perform new task–agent combinations or serve as a starting point for good initializations when learning complex behaviors. Many multi-task problems also highlight the issue of parameter sharing between distinct tasks. To that end, Yang et al. (2020) introduces a base policy network composed of multiple modules and a separate routing network. The routing network uses a task embedding and the current state of the agent to reconfigure the base network's modules with a learned routing strategy.

In contrast, our Active Dendrites Network activates sparse subnetworks by introducing control over individual neurons in a network. By dynamically integrating a context vector to modulate these neurons, the network automatically creates distinct subnetworks to learn each task. Unlike prior approaches, our network does not require modified learning rules, separate modules, or dedicated routing networks to train new tasks. Rather, a single architecture is capable of reducing gradient interference, learning a diverse range of tasks, and can be applied to scenarios beyond multi-task RL.

## 5.3. Comparison to Other Continual Learning Systems

There are a few papers on continual learning that are very related to the core ideas in this paper. Our networks create representations composed of different sparse subnetworks of neurons. Abbasi et al. (2022) use $k$WTA in conjunction with a modified gradient update method to avoid task interference. XdG (Masse et al., 2018) and Supermasks (Wortsman et al., 2020) also explicitly utilize sparse subnetworks per task. XdG, discussed extensively in Section 4.2.4, hard-codes a sparse subnetwork for each task. This extra supervision step removes the need to dynamically gate activations but requires knowledge of the task identity during inference. In addition, as seen in **Figure 9**, XdG does not scale as well as our networks. In contrast, Supermasks uses a randomly initialized network and focuses on locating the best subnetwork for each task and forgoes any further training. The technique shows impressive scaling behavior, but it's unclear whether complex tasks can be solved without any network training.

Our Active Dendrites Neurons dynamically determine a representation for each feedforward input based on auxiliary contextual inputs. In the case where the modulation function $f$ involves multiplication, our Active Dendrites Networks are an instance of multiplicative networks. Jayakumar et al. (2020) demonstrated that multiplicative networks can excel in multi-task scenarios by learning dynamic representations in a task-specific manner.

Several ANN-based techniques leverage the idea of auxiliary contextual inputs. For instance, Gated Linear Networks (Veness et al., 2021) and Dendritic Gated Networks (Sezener et al., 2021) gate activation values for each neuron based on contextual information. Although inspired by dendrites these models (1) don't activate sparse subnetworks, (2) have fixed random dendritic weights (to model cerebellar dendritic branches), and (3) are binary classifiers (i.e., 10 Dendritic Gated Networks are required to classify MNIST digits). Furthermore, because Sezener et al. (2021) test Dendritic Gated Networks only up to 10

permutedMNIST tasks using a very different metric, we cannot provide a direct comparison with our model.

## 5.4. Future Work

Our initial results show that active dendrites and sparse representations can mitigate catastrophic forgetting and interference in multi-task RL and continual learning settings. One crucial next step is to test this framework on more real-world scenarios with greater complexity than MT10 or permutedMNIST. The majority of existing work in MTRL considers tasks with shared input and action spaces. Dendrites may be beneficial in scenarios where this assumption does not hold. Extending to tasks with very different input and output spaces is an interesting area for future research. Another interesting area is to combine our two scenarios and explore continual multi-task RL. While testing on more diverse benchmarks, it will also be important to explore additional methods for generating context vectors for a given task. Another important direction for future research is to investigate sparse dendritic segments, following neuroscience evidence suggesting that each segment relies on just a handful of synapses (Branco and Häusser, 2011).

# 6. METHODS

## 6.1. Multi-Task Reinforcement Learning Experiments

In this section, we provide the details of our multi-task RL experiments[3]. We use the Multi-Task Soft Actor-Critic algorithm (MTSAC) originally discussed in Yu et al. (2019), which is described as an adaptation of the Soft Actor-Critic algorithm (SAC; Haarnoja et al., 2018b). We adapt the code in the original Meta-World GitHub repository[4] to fit our experiments.

### 6.1.1. Basics of Reinforcement Learning

To formalize our specific RL problem, we define some fundamental concepts. The state of the RL agent and the action it will take at a specific time $t$ are denoted as $s_t$ and $a_t$, respectively. The RL algorithm trains a policy $\pi$ to take $a_t$ given $s_t$ in order to maximize total return $G = \sum_t \gamma^t r(a_t, s_t)$ across all time-steps $t$, where $r(a_t, s_t)$ is the reward given by the environment and $\gamma$ is a discount factor to strongly consider immediate rewards.

To optimize this policy, our RL formulation uses Markov Decision Processes (MDPs) to model decision making in stochastic environments. Following the notation introduced in Sutton and Barto (2018), we consider a finite-horizon MDP defined by the tuple $(S, A, P, r, T)$ that operates in a state space $S$ and action space $A$. The MDP also uses the transition probability $P$ between any two states $s_t$ and $s_{t+1}$ by taking action $a_t$, which is explicitly defined across all states and actions as $P(s_{t+1}|s_t, a_t) : S \times A \to \mathbb{R}$. Agents in this setting receive a reward $r : S \times A \to \mathbb{R}$ that is also defined across all states and actions. Additionally, agents must make decisions within a fixed number of steps, denoted by the finite-time horizon $T$.

The RL algorithm we consider computes a value function that estimates the total return accrued at a specific state. More precisely, the value function describes the significance of starting at some state $s_t$ and following some policy $\pi$. The value function for policy $\pi$ can be defined below:

$$V_\pi(s_t) = \mathbb{E}_{a_t \sim \pi}\left[r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim P}\left[\gamma V_\pi(s_{t+1})\right]\right] \tag{7}$$

$$= \mathbb{E}_{a_t \sim \pi}\left[r(s_t, a_t) + \sum_{s_{t+1} \in S} P(s_{t+1}|s_t, a_t)\left(\gamma V_\pi(s_{t+1})\right)\right] \tag{8}$$

$$= \sum_{a_t \in A} \pi(a_t|s_t)\left[r(s_t, a_t) + \sum_{s_{t+1} \in S} P(s_{t+1}|s_t, a_t)\left(\gamma V_\pi(s_{t+1})\right)\right] \tag{9}$$

Note that Equation (7) establishes a recursive relation with respect to the function $V_\pi$. To estimate the value at a given state $s_t$, an agent must take an action $a_t$ sampled from policy $\pi$ to calculate the expected value at the next state $s_{t+1}$. By repeating this process until a terminal state is reached, the agent can use the value function to choose actions that lead to highly valued states.

The RL algorithm we consider also estimates an action-value function $Q_\pi$. While value functions estimate the value of starting at $s_t$ and following $\pi$, action-value functions estimate the value of starting at $s_t$, taking action $a_t$, and *then* following $\pi$ until a terminal state is reached. This is known as the $Q$ function, which can be described explicitly below:

$$Q_\pi(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim P}\left[\gamma \mathbb{E}_{a_{t+1} \sim \pi}\left[Q_\pi(s_{t+1}, a_{t+1})\right]\right] \tag{10}$$

Fundamentally, value functions and $Q$ functions can be related by the following two expressions:

$$Q_\pi(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim P}\left[\gamma V_\pi(s_{t+1})\right] \tag{11}$$

$$V_\pi(s_t) = \mathbb{E}_{a_t \sim \pi}\left[Q_\pi(s_t, a_t)\right] \tag{12}$$

Throughout the training process, explored state, action, reward, and next state transitions–namely $(s_t,\ a_t,\ r_t,\ s_{t+1})$–are used to train $\pi$. In some algorithms, including ours, these transitions are stored in a replay buffer $\mathcal{D}$ and are sampled by batch during each step of training to dynamically compute either the value or $Q$ function. After a suitable period of exploration, the agent will take actions that yield the maximum $V_\pi(s_t)$ or $Q_\pi(s_t, a_t)$ value. Note that while $V$, $Q$, and $\pi$ are expressed in discrete state and action spaces above, they can be easily extended to work in continuous state and action spaces using function approximations such as neural networks.

### 6.1.2. Basics of Multi-Task Reinforcement Learning

We can extend the ideas in Section 6.1.1 to our multi-task RL experiments in Meta-World. Specifically, the problem framework uses a separate MDP to model each task $\tau$. In the context of the Meta-World multi-task environment, each task shares

[3]PyTorch source code for our experiments is available at https://github.com/numenta/htmpapers.

[4]Meta-World source code is available at https://github.com/rlworkgroup/metaworld.

**TABLE 2 |** The hyperparameters for each multi-task RL model.

| | Active dendrites network | MLP baseline | Large MLP |
|---|---|---|---|
| **Network hyperparameters** | | | |
| Feedforward input size | 39 | 49 | 49 |
| Hidden sizes | 2 × [2,800] | 2 × [2,800] | 2 × [3,000] |
| Output size | 4 | 4 | 4 |
| Feedforward weight sparsity | 10% | 0% | 0% |
| Activation function | *k*WTA | ReLU | ReLU |
| Activation sparsity | 25% | ∼ 50% | ∼ 50% |
| Num. dendritic segments per neuron | 10 | 0 | 0 |
| Num. weights per dendritic segment | 10 | — | — |
| Num. hidden layers modulated | 1 | — | — |
| Dendritic segment weight sparsity | 0% | — | — |
| Non-zero feedforward parameters | 7,169,964 | 7,994,004 | 9,165,004 |
| Non-zero dendritic parameters | 280,000 | 0 | 0 |
| Non-zero parameters (total) | 7,449,964 | 7,994,004 | 9,165,004 |
| **Training hyperparameters** | | | |
| Number of epochs | 3,000 | 3,000 | 3,000 |
| Number of timesteps | 15,000,000 | 15,000,000 | 15,000,000 |
| Number of gradient steps per epoch | 250 | 250 | 250 |
| Buffer sampling batch size | 2,560 | 2,560 | 2,560 |
| Replay buffer size | 1,000,000 | 1,000,000 | 1,000,000 |
| Policy learning rate | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ |
| Q-function learning rate | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ |
| Target Q-function update rate | $5 \times 10^{-3}$ | $5 \times 10^{-3}$ | $5 \times 10^{-3}$ |
| Policy (Min, Max) Std | $e^{-20}$, $e^{2}$ | $e^{-20}$, $e^{2}$ | $e^{-20}$, $e^{2}$ |
| Action sampling distribution type | Tanh normal | Tanh normal | Tanh normal |

identical state and actions spaces and defines common transition probabilities and time horizons. However, each task defines separate reward functions, although all functions share a similar scale and structure to allow a single agent to uniformly learn all tasks. We assume a uniform distribution of tasks $p(\tau)$ and train a task-conditioned, stochastic policy $\pi(a|s, \boldsymbol{c})$ to solve all $\mathcal{T}$ tasks, where $\boldsymbol{c}$ is a context vector that provides information about a specific task. Explicitly, the policy is trained to maximize the total return from the task distribution $p(\tau)$ as expressed by $\mathbb{E}_{\tau \sim p(\tau)} \left[ E_{\pi} \left[ \Sigma_{t=0}^{T} \gamma^{t} r_t(s_t, a_t) \right] \right]$.

### 6.1.3. The Multi-Task Soft-Actor Critic Algorithm

The MTSAC algorithm we use in our experiments is based on the SAC algorithm and slightly modified to solve $\tau$ various tasks simultaneously. In SAC, an RL algorithm uses a $V$ or $Q$ network (known as the critic) to train a policy $\pi$ (known as the actor) to take better actions. SAC modifies the original value function definition to also consider the entropy of the policy $\pi$. By maximizing *both* expected return and entropy, an agent is motivated to explore new states while computing an optimal policy. More details about the SAC algorithm can be found in Haarnoja et al. (2018b).

In MTSAC, both $\pi$ and $Q$ are conditioned by context vector $\boldsymbol{c}$ and are thus denoted as $\pi(a_t|s_t, \boldsymbol{c})$ and $Q(s_t, a_t|\boldsymbol{c})$, respectively. MTSAC also uses $\tau$ different entropy coefficients $\alpha_\tau$ to control the exploration per task. More details about the MTSAC algorithm can be found in Yu et al. (2019).

The Meta-World environment we use is MT10, which contains 10 different tasks that a single robotic arm must solve. All tasks share an identical state space $s_t \in \mathbb{R}^{39}$ and action space $a_t \in \mathbb{R}^{4}$. Because there are 10 different tasks, $\boldsymbol{c}$ is a 10-dimensional one-hot encoded vector that describes the task ID.

### 6.1.4. Experiment Settings

The training hyperparameters are identical for the Active Dendrites Network and the MLP baseline. For every run, the model is trained for 3,000 epochs. Each epoch comprises of one episode of 500 timesteps for each of the 10 tasks. In total, this amounts to 5,000 timesteps per epoch and 15,000,000 timesteps for the entire run. Our implementation parallels the baseline implementation. In our experiments, we use one Active Dendrites Network to model the policy $\pi$ and another to model the $Q$ function.

The model is also used to collect new data to be stored in a replay buffer. At the end of each epoch, the model is then trained for 250 gradient steps. For each gradient step, the algorithm randomly samples a batch of 2,560 experiences from the replay buffer. The replay buffer is a queue of limited size, capped at 1 million, with newer experiences replacing older experiences.

To allow a better comparison between the models, we set the learning rates, target $Q$ function update rate, and policy

minimum and maximum standard deviations to be the same for all runs. The difference between the models is the network architecture of the policy and Q functions. The Active Dendrites Network modulates each neuron in the second hidden layer with 10 dendritic segments, where each segment is a vector of size 10. In total, this dendritic layer adds an additional 280,000 parameters to the overall network. We apply a fixed sparsity mask of 10% to the weights of the feedforward layers of the Active Dendrites Network to reduce the number of free parameters and keep it of comparable size to the MLP baseline. The Active Dendrites Network also uses a $k$WTA activation function instead of ReLU, which effectively selects the top 25% of units and zeroes out the remaining during every forward step.

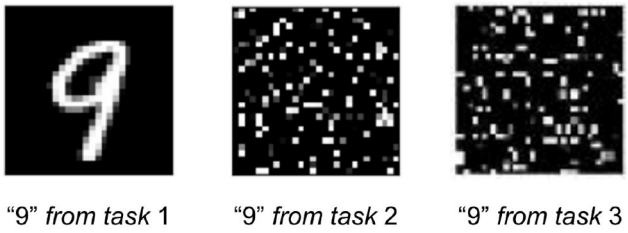



**FIGURE 15 |** A visual illustration of permutedMNIST. Each task applies a unique pixel-wise permutation to the same original image (leftmost image) while preserving the target label. A model's task is to identify the digit in each case regardless of permutation.

## 6.2. Continual Learning Experiments

We discuss the setup of the continual learning experiments. Our model is trained on $\mathcal{T}$ discrete tasks in sequence. More specifically, our model first trains on task $\tau = 1$. Once learning task $\tau$ is complete, the model then starts training on task $\tau + 1$. After training on task $\tau = \mathcal{T}$, all learning is complete. Each task $\tau$ consists of standard batch learning with i.i.d. training data. While training on task $\tau$ where $1 \leq \tau \leq \mathcal{T}$, our model only receives training data corresponding to task $\tau$. Once the model finishes learning task $\tau$, it never again receives information about any task $\tau' \leq \tau$ for training purposes. The model is, however, evaluated on the test data for each task to determine how well it performs.

### 6.2.1. PermutedMNIST

We train our model on the permutedMNIST dataset, a benchmark dataset for continual learning (Goodfellow et al., 2014), which is derived from MNIST. MNIST comprises approximately 60,000 black and white images of handwritten digits 0–9 where each such image has dimensions $28 \times 28$ pixels and the associated target digit as the label. During training, roughly 50,000 images are used for training and the remaining 10,000 for testing.

In permutedMNIST with $\mathcal{T}$ tasks, MNIST is replicated $\mathcal{T}$ times, but each time with a unique pixel-wise permutation applied to all 60,000 images. That is, each task randomly re-arranges the pixels of all images exactly the same way while preserving the associated target label. The first task ($\tau = 1$) corresponds to the identity permutation (i.e., regular MNIST) and every subsequent task generates a random pixel-wise permutation. As permutedMNIST is synthesized from regular MNIST, there can be an arbitrary number of tasks, $\mathcal{T}$. **Figure 15** illustrates a single image taken from different tasks.

Our model, and all comparisons we made, uses a single output head. Each model has 10 output units in the final layer of the network representing the 10 categories. These output units are re-used for each task, i.e., the model is trained to predict the first output unit for label "0" regardless of which task the input data corresponds to. In this setup chance accuracy is 10%.

### 6.2.2. Experiment Settings

When employing the prototype method described in Section 4.2.1 to select context signals at test time only, we train an Active Dendrites Network with two hidden layers that comprise Active Dendrites Neurons. We find that having just a single hidden layer reduced accuracy by a few percentage points while 3 hidden layers provided a minimal performance boost. For 100 tasks, a single layer reduced accuracy by 3% and three layers improved accuracy by 0.5%. For all training, we use the Adam optimizer (Kingma and Ba, 2015) and a batch size of 256 samples. **Table 3** gives the exact hyperparameters and model architecture for each model we train and evaluate on permutedMNIST. Note that hyperparameters were optimized individually for each setting.

To combine Active Dendrites Network with SI, and to compare against XdG, we reduce the number of units in each hidden layer from 2,048 to 2,000 as to exactly match the architectures (with the exception of dendritic segments) used in the SI and XdG papers. (See **Supplementary Materials** for a discussion on the number of parameters.) In addition, the SI-and-Active-Dendrites network is trained for 20 epochs per task instead of just three as this significantly improves results. We fix the learning rate to be $5 \times 10^{-4}$ for all numbers of tasks, and we use SI regularization strength $c = 0.1$ and damping coefficient $\xi = 0.1$. Both (a) training for 20 epochs per task and (b) the $c, \xi$ values that we use here align with the training setups of Zenke et al. (2017) and Masse et al. (2018).

### 6.2.3. Constructing Prototypes During Training Without Task Information

When task information is not given during training nor testing, the task corresponding to each input example must be inferred. This section describes the online clustering method we implemented to infer task information during training. One inductive bias in our procedure is that all training examples in a batch correspond to the same task, since continual learning scenarios usually only observe examples from a single task within a given batch.

Formally, let $X = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$ be a batch of $n$ training examples (in the case of permutedMNIST, each $\boldsymbol{x}^{(i)}$ is a 784-dimensional vector for $1 \leq i \leq n$). Suppose $M$ individual prototypes are designated thus far: $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_M$. For each $\boldsymbol{p}_j$ (where $1 \leq j \leq M$), the individual examples used to construct that prototype are also stored in memory: $Y_j = \{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m_j)}\}$, where $m_j$ gives the number of examples for cluster $j$. These

**TABLE 3 |** The hyperparameters used to train each model on permutedMNIST.

| | Active dendrites network (task info. provided) | Active dendrites network (task info. not provided) | 3-Layer MLP | 10-Layer MLP |
|---|---|---|---|---|
| **Network hyperparameters** | | | | |
| Feedforward input size | 784 | 784 | 784 | 784 |
| Hidden sizes | 2 × [2,048] | 2 × [2,048] | 2 × [2,048] | 10 × [2,048] |
| Output size | 10 | 10 | 10 | 10 |
| Feedforward weight sparsity | 50% | 50% | 0% | 0% |
| Activation function | *k*WTA | *k*WTA | ReLU | ReLU |
| Activation sparsity | 5% | 5% | ∼ 50% | ∼ 50% |
| Num. dendritic segments per neuron | $\mathcal{T}$ | $\mathcal{T}$ | 0 | 0 |
| Num. weights per dendritic segment | 784 | 784 | – | – |
| Dendritic segment weight sparsity | 0% | 0% | – | – |
| Non-zero feedforward parameters | 2,914,314 | 2,914,314 | 5,824,522 | 35,198,986 |
| Non-zero dendritic parameters | $\mathcal{T} \times 3,211,264$ | $\mathcal{T} \times 3,211,264$ | 0 | 0 |
| Non-zero parameters (total) | See **Supplementary Materials** | | 5,824,522 | 35,198,986 |
| **Training hyperparameters** | | | | |
| ($\mathcal{T} = 2$) Learning rate | $5 \times 10^{-4}$ | $10^{-3}$ | – | – |
| ($\mathcal{T} = 2$) Number of epochs | 1 | 5 | – | – |
| ($\mathcal{T} = 5$) Learning rate | $5 \times 10^{-4}$ | $10^{-3}$ | – | – |
| ($\mathcal{T} = 5$) Number of epochs | 1 | 5 | – | – |
| ($\mathcal{T} = 10$) Learning rate | $5 \times 10^{-4}$ | $10^{-3}$ | $3 \times 10^{-6}$ | $3 \times 10^{-6}$ |
| ($\mathcal{T} = 10$) Number of epochs | 3 | 3 | 5 | 3 |
| ($\mathcal{T} = 25$) Learning rate | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ | – | – |
| ($\mathcal{T} = 25$) Number of epochs | 5 | 1 | – | – |
| ($\mathcal{T} = 50$) Learning rate | $3 \times 10^{-4}$ | $10^{-4}$ | – | – |
| ($\mathcal{T} = 50$) Number of epochs | 3 | 3 | – | – |
| ($\mathcal{T} = 100$) Learning rate | $10^{-4}$ | $10^{-4}$ | $10^{-6}$ | $3 \times 10^{-7}$ |
| ($\mathcal{T} = 100$) Number of epochs | 3 | 3 | 3 | 3 |

previous training examples are observed by the learner during previous batches of learning and stored in memory. We identify if the new batch $X$ is similar enough to any cluster of training examples $Y_j$ such that the corresponding prototype $\boldsymbol{p}_j$ should be used as the context signal. If a cluster $j$ is found such that $X$ is "similar" to $Y_j$, then $Y_j$ is expanded to include $X$. Subsequently, $\boldsymbol{p}_j$ is updated to incorporate samples from $X$. Otherwise, if $X$ is deemed significantly different from $Y_j$ for all $j$, then a new cluster is formed: $Y_{M+1} \leftarrow X$ and its prototype is the element-wise mean of all $\boldsymbol{x} \in X$. **Algorithm 1** describes the procedure for clustering during training when task information is not provided.

In the pseudocode, how do we determine when $X$ is similar enough to some $Y_j$? If we have univariate data (i.e., if each $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y_j$ is a scalar quantity), we could use an unpaired $t$-test do this. Instead, we use a generalized version of an unpaired $t$-test that applies to multivariate data. In our hypothesis testing setup, the null hypothesis is that for any given $j$, the same underlying process generates samples from both $X$ and $Y_j$. When we accept the null hypothesis, we assume each $\boldsymbol{x} \in X$ and each $\boldsymbol{y} \in Y_j$ are training examples from the same permutedMNIST task—and therefore $\boldsymbol{p}_j$ can be used as the context signal when training an Active Dendrites Network on examples in $X$ (albeit $\boldsymbol{p}_j$ is first updated to account for $X$).

**Algorithm 1** : Clustering algorithm by which a new batch of inputs $X$ either gets assigned to one of $M$ existing clusters or initiates cluster $M+1$. This procedure is greedy since it assigns $X$ to the first cluster $j$ that it suitably matches.

**procedure** CLUSTER($X$, $Y$)
  $M \leftarrow 0$ ▷ Number of existing clusters
  **while** not done learning **do**
    $X \leftarrow$ new batch
    assigned $\leftarrow$ **False**
    **for** $j = 1$ to $M$ **do**
      **if** ¬ assigned and IS_MATCH($X$, $Y_j$) **then**
        assigned $\leftarrow$ **True**
        $Y_j \leftarrow Y_j \cup X$
        update $\boldsymbol{p}_j$ to include each $\boldsymbol{x} \in X$

Hotelling (1931) proposed Hotelling's $t$-squared statistic ($t^2$) as a generalization of the $t$-statistic used to perform single-variable $t$-tests; it is computed as

$$t^2 = \frac{|X||Y_j|}{|X| + |Y_j|} \left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right)$$

where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are simply the element-wise means of all $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y_j$, respectively, and $\boldsymbol{\Sigma}$ is the pooled, sample-adjusted covariance matrix of samples in $X$ and $Y_j$. The test statistic $t^2$ can be compared to a chosen $p$-value to accept or reject the null hypothesis by first transforming it to a value drawn from an $F$-distribution (whose cumulative density function is more well-studied than that of the $t$-squared distribution) as follows:

$$f = \frac{|X| + |Y_j| - d - 1}{d\left(|X| + |Y_j| - 2\right)} t^2$$

where $d$ is the dimensionality of the samples.

We fix a $p$-value and derive a value for $f$ based on $t^2$ as give above. If $f > p$, then we reject the null hypothesis since the probability that the same generative process explains both $X$ and $Y_j$ is extremely low, and thus create a new cluster. Since we perform pairwise multivariate $t$-tests between $X$ and $Y_j$ for all existing prototypes $j$, a new cluster and prototype emerge if and only if we reject the null hypothesis for all $M$ $t$-tests. **Algorithm 2** describes the procedure for performing the multivariate $t$-test *via* the $t$-squared statistic given two sets of multivariate samples.

**Algorithm 2** : Unpaired multivariate $t$-test using Hotelling's $t$-squared statistic. Here, we use a slight abuse of notation when computing covariance matrices by assuming sets of $d$-dimensional vectors can also be treated as matrices whose rows correspond to their $d$-dimensional elements. We assume a $p$-value is fixed a priori. In our implementation, we replace all standard matrix inversions with the Moore-Penrose pseudo-inversion.

**procedure** IS_MATCH($X$, $Y$)
  $\bar{\boldsymbol{x}} \leftarrow \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} \boldsymbol{x}$ ▷ Compute $X$ mean
  $\bar{\boldsymbol{y}} \leftarrow \frac{1}{|Y|} \sum_{\boldsymbol{y} \in Y} \boldsymbol{y}$ ▷ Compute $Y$ mean
  $\boldsymbol{\Sigma}_X \leftarrow \frac{1}{|X|-1}(X - \bar{\boldsymbol{x}})(X - \bar{\boldsymbol{x}})^\top$ ▷ Compute $X$ covariance
  $\boldsymbol{\Sigma}_Y \leftarrow \frac{1}{|Y|-1}(X - \bar{\boldsymbol{y}})(Y - \bar{\boldsymbol{y}})^\top$ ▷ Compute $Y$ covariance
  $\boldsymbol{\Sigma} \leftarrow \frac{(|X|-1)\boldsymbol{\Sigma}_X + (|Y|-1)\boldsymbol{\Sigma}_Y}{|X|+|Y|-2}$ ▷ Compute pooled covariance
  $t^2 \leftarrow \frac{|X||Y_j|}{|X|+|Y_j|}\left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right)$ ▷ Compute $t^2$
  $f = \frac{|X|+|Y_j|-d-1}{d(|X|+|Y_j|-2)} t^2$ ▷ Convert $t^2$ to $f$
  **if** $f > p$ **then**:
    **return False** ▷ Reject null hypothesis
  **else**
    **return True** ▷ Accept null hypothesis

## 6.3. Absolute Max Gating

We outline how we implement gating in Active Dendrites Networks. In Section 3, we present gating as modifying the value of the weighted linear sum computed by the point neuron based on the maximum activation, i.e., $\sigma(\max_j \boldsymbol{u}^\top \boldsymbol{c})$. One problem with this formulation is that it becomes difficult to turn a neuron off (i.e., force it's activation value to be zero) due to the max operator. That is, if dendritic segment $j$ learns to turn off the unit, then based on sigmoidal gating, we should expect that $\boldsymbol{u}_j^\top \boldsymbol{c}$ is a small number with large absolute value (very negative). However, it's likely that for some other segment $j'$ ($j \neq j'$), $\boldsymbol{u}_{j'}^\top \boldsymbol{c} > 0 > \boldsymbol{u}_j^\top \boldsymbol{c}$ which means that segment $j'$ will be selected by the max operator instead of segment $j$, hence increasing the chance that the neuron will be selected by the $k$WTA process.

This motivates absolute max gating in which the activation with the largest magnitude is selected and its sign is kept. More formally, a point neuron augmented with absolute max gating computes its output as

$$j^* = \arg\max_j \left|\boldsymbol{u}_j^\top \boldsymbol{c}\right|,$$
$$\hat{y} = \left(\boldsymbol{w}^\top \boldsymbol{x} + b\right) \sigma\left(\boldsymbol{u}_{j^*}^\top \boldsymbol{c}\right).$$

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

SA conceived of the overall theory and the mapping to neuroscience. KG and SA implemented the core dendrites code. KG and JF implemented the continual learning experiments and their analysis. AV originally conceived of the mapping to multi-task RL and designed the original RL experiments. AI, LS, and AV implemented and ran the multi-task RL experiments. SA, AI, AV, and KG contributed to the article design and wrote the text. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

We thank Jeff Hawkins, Ben Cohen, Greg Maltz, Luiz Scheinkman, and the rest of the Numenta team for helpful feedback throughout the editing process.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2022.846219/full#supplementary-material

## REFERENCES

Abbasi, A., Nooralinejad, P., Braverman, V., Pirsiavash, H., and Kolouri, S. (2022). Sparsity and heterogeneous dropout for continual learning in the null space of neural activations. *arXiv:2203.06514*. doi: 10.48550/arXiv.2203.06514

Ahmad, S., and Hawkins, J. (2016). How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites. *arXiv:1601.00720*. doi: 10.48550/arXiv.1601.00720

Ahmad, S., and Scheinkman, L. (2019). How can we be so dense? The benefits of using highly sparse

representations. *arXiv:1903.11257*. doi: 10.48550/arXiv.1903.11257

Andreas, J., Klein, D., and Levine, S. (2017). "Modular multitask reinforcement learning with policy sketches," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney, NSW).

Antic, S. D., Hines, M., and Lytton, W. W. (2018). Embedded ensemble encoding hypothesis: The role of the "prepared" cell. *J. Neurosci. Res*. 96, 1543–1559. doi: 10.1002/jnr.24240

Antic, S. D., Zhou, W.-L., Moore, A. R., Short, S. M., and Ikonomu, K. D. (2010). The decade of the dendritic NMDA spike. *J. Neurosci. Res*. 88, 2991–3001. doi: 10.1002/jnr.22444

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). A brief survey of deep reinforcement learning. *IEEE Signal Process. Mag*. 34, 26–38. doi: 10.1109/MSP.2017.2743240

Attwell, D., and Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab*. 21, 1133–1145. doi: 10.1097/00004647-200110000-00001

Barth, A. L., and Poulet, J. F. a. (2012). Experimental evidence for sparse firing in the neocortex. *Trends Neurosci*. 35, 345–355. doi: 10.1016/j.tins.2012.03.008

Beniaguev, D., Segev, I., and London, M. (2021). Single cortical neurons as deep artificial neural networks. *Neuron* 109, 2727–2739. doi: 10.1016/j.neuron.2021.07.002

Bentivoglio, M., and Swanson, L. W. (2001). On the fine structure of the PES hippocampi major. *Brain Res. Bull*. 54, 461–483. doi: 10.1016/S0361-9230(01)00430-0

Branco, T., and Häusser, M. (2010). The single dendritic branch as a fundamental functional unit in the nervous system. *Curr. Opin. Neurobiol*. 20, 494–502. doi: 10.1016/j.conb.2010.07.009

Branco, T., and Häusser, M. (2011). Synaptic integration gradients in single cortical pyramidal cell dendrites. *Neuron* 69, 885–892. doi: 10.1016/j.neuron.2011.02.006

Caruana, R. (1997). Multitask learning. *Mach. Learn*. 28, 41–75. doi: 10.1023/A:1007379606734

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. (2018). "Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning* (Stockholm).

Cui, Y., Ahmad, S., and Hawkins, J. (2017). The HTM Spatial Pooler–a neocortical algorithm for online sparse distributed coding. *Front. Comput. Neurosci*. 11:111. doi: 10.3389/fncom.2017.00111

Devin, C., Gupta, A., Darrell, T., Abbeel, P., and Levine, S. (2017). "Learning modular neural network policies for multi-task and multi-robot transfer," in *Proceedings of the IEEE International Conference on Robotics and Automation* (Singapore). doi: 10.1109/ICRA.2017.7989250

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). "Multi-task learning for multiple language translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Beijing: Association for Computational Linguistics), 1723–1732. doi: 10.3115/v1/P15-1166

Du, Y., Czarnecki, W. M., Jayakumar, S. M., Farajtabar, M., Pascanu, R., and Lakshminarayanan, B. (2020). Adapting auxiliary losses using gradient similarity. *arXiv:1806.08730*. doi: 10.48550/arXiv.1812.02224

Errington, A. C., and Connelly, W. M. (2011). Dendritic T-type Ca2+ channels: giving a boost to thalamic reticular neurons. *J. Neurosci*. 31, 5551–5553. doi: 10.1523/JNEUROSCI.0067-11.2011

Flesch, T., Nagy, D. G., Saxe, A., and Summerfield, C. (2022). Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *arXiv [Preprint]. arXiv: 2203.11560*. Available online at: https://arxiv.org/ftp/arxiv/papers/2203/2203.11560.pdf

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci*. 3, 128–135. doi: 10.1016/S1364-6613(99)01294-2

Gao, P. P., Graham, J. W., Zhou, W. L., Jang, J., Angulo, S., Dura-Bernal, S., et al. (2021). Local glutamate-mediated dendritic plateau potentials change the state of the cortical pyramidal neuron. *J. Neurophysiol*. 125, 23–42. doi: 10.1152/jn.00734.2019

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. (2014). "An empirical investigation of catastrophic forgetting in gradient-based neural networks," in *Proceedings of the 2nd International Conference on Learning Representations* (Banff, AB).

Goyal, A., Sodhani, S., Binas, J., Peng, X. B., Levine, S., and Bengio, Y. (2020). "Reinforcement learning with competitive ensembles of information-constrained primitives," in *Proceedings of the 8th International Conference on Learning Representations, Digital*.

Guest, J. M., Bast, A., Narayanan, R. T., and Oberlaender, M. (2021). Thalamus gates active dendritic computations in cortex during sensory processing. *bioRxiv*. doi: 10.1101/2021.10.21.465325

Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., and Levine, S. (2018a). "Composable deep reinforcement learning for robotic manipulation," in *Proceedings of the IEEE International Conference on Robotics and Automation* (Brisbane, QLD). doi: 10.1109/ICRA.2018.8460756

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018b). "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning* (Stockholm).

Hawkins, J., and Ahmad, S. (2016). Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Front. Neural Circuits* 10:23. doi: 10.3389/fncir.2016.00023

Hawkins, J., Ahmad, S., and Cui, Y. (2017). A theory of how columns in the neocortex enable learning the structure of the world. *Front. Neural Circuits* 11:81. doi: 10.3389/fncir.2017.00081

Heald, J. B., Lengyel, M., and Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature* 600, 489–493. doi: 10.1038/s41586-021-04129-3

Holmgren, C., Harkany, T., Svennenfors, B., and Zilberter, Y. (2003). Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *J. Physiol*. 551, 139–153. doi: 10.1113/jphysiol.2003.044784

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat*. 2, 360–378. doi: 10.1214/aoms/1177732979

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., and Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. *Int. J. Robot. Res*. 40, 698–721. doi: 10.1177/0278364920987859

Irpan, A. (2018). *Deep Reinforcement Learning Doesn't Work Yet*. Available online at: https://www.alexirpan.com/2018/02/14/rl-hard.html (accessed November 1, 2021).

Jadi, M. P., Behabadi, B. F., Poleg-Polsky, A., Schiller, J., and Mel, B. W. (2014). An augmented two-layer model captures nonlinear analog spatial integration effects in pyramidal neuron dendrites. *Proc. IEEE* 102, 782–798. doi: 10.1109/JPROC.2014.2312671

Jayakumar, S. M., Czarnecki, W. M., Menick, J., Schwarz, J., Rae, J., Osindero, S., et al. (2020). "Multiplicative interactions and where to find them," in *Proceedings of the 8th International Conference on Learning Representations, Digital*.

Kandel, E. (2012). *Principles of Neural Science, 5th Edn*. New York, NY: McGraw-Hill.

Kendall, A., Gal, Y., and Cipolla, R. (2019). "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA).

Kerlin, A., Boaz, M., Flickinger, D., Maclennan, B. J., Dean, M. B., Davis, C., et al. (2019). Functional clustering of dendritic activity during decision-making. *eLife* 8:e46966. doi: 10.7554/eLife.46966

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations* (San Diego, CA).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U.S.A*. 114, 3521–3526. doi: 10.1073/pnas.1611835114

Lafourcade, M., van der Goes, M.-S. H., Vardalaki, D., Brown, N. J., Voigts, J., Yun, D. H., et al. (2022). Differential dendritic integration of long-range inputs in association cortex *via* subcellular changes in synaptic AMPA-to-NMDA receptor ratio. *Neuron*. doi: 10.1016/j.neuron.2022.01.025

Lapique, L. (1907). Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. *J. Physiol. Pathol*. 9, 620–635.

Larkum, M. (2022). Are dendrites conceptually useful? *Neurosci*. S0306-4522(22)00120-8. doi: 10.1016/j.neuroscience.2022.03.008. [Epub ahead of print].

Larkum, M. E., Zhu, J. J., and Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* 398, 338–341. doi: 10.1038/18686

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Liang, F., Li, H., Chou, X.-l., Zhou, M., Zhang, N. K., Xiao, Z., et al. (2019). Sparse representation in awake auditory cortex: cell-type dependence, synaptic mechanisms, developmental emergence, and modulation. *Cereb. Cortex* 29, 3796–3812. doi: 10.1093/cercor/bhy260

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2016). "Continuous control with deep reinforcement learning," in *Proceedings of the 4th International Conference on Learning Representations* (San Juan).

Limbacher, T., and Legenstein, R. (2020). Emergence of stable synaptic clusters on dendrites through synaptic rewiring. *Front. Comput. Neurosci*. 14:57. doi: 10.3389/fncom.2020.00057

Liu, S., Johns, E., and Davison, A. J. (2019). "End-to-end multi-task learning with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA). doi: 10.1109/CVPR.2019.00197

London, M., and Häusser, M. (2005). Dendritic computation. *Annu. Rev. Neurosci*. 28, 503–532. doi: 10.1146/annurev.neuro.28.061604.135703

Losonczy, A., Makara, J. K., and Magee, J. C. (2008). Compartmentalized dendritic plasticity and input feature storage in neurons. *Nature* 452, 436–441. doi: 10.1038/nature06725

Magee, J. C. (2000). Dendritic integration of excitatory synaptic input. *Nat. Rev. Neurosci*. 1, 181–190. doi: 10.1038/35044552

Majani, E., Erlanson, R., and Abu-Mostafa, Y. (1989). "On the k-winners-take-all network," in *Advances in Neural Information Processing Systems* (Denver, CO).

Major, G., Larkum, M. E., and Schiller, J. (2013). Active properties of neocortical pyramidal neuron dendrites. *Annu. Rev. Neurosci*. 36, 1–24. doi: 10.1146/annurev-neuro-062111-150343

Maninis, K.-K., Radosavovic, I., and Kokkinos, I. (2019). "Attentive single-tasking of multiple tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA). doi: 10.1109/CVPR.2019.00195

Masse, N. Y., Grant, G. D., and Freedman, D. J. (2018). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proc. Natl. Acad. Sci. U.S.A*. 115, E10467–E10475. doi: 10.1073/pnas.1803839115

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv:1806.08730*. doi: 10.48550/arXiv.1806.08730

McClelland, J. L., Rumelhart, D. E., and the PDP Research Group (1986). *Parallel Distributed Processing, 2nd Edn*. Cambridge, MA: MIT Press.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv*. 24, 109–165. doi: 10.1016/S0079-7421(08)60536-8

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV). doi: 10.1109/CVPR.2016.433

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., et al. (2013). "Playing atari with deep reinforcement learning," in *Advances in Neural Information Processing Systems* (Lake Tahoe).

Paiton, D. M., Frye, C. G., Lundquist, S. Y., Bowen, J. D., Zarcone, R., and Olshausen, B. A. (2020). Selectivity and robustness of sparse coding networks. *J. Vis*. 20:10. doi: 10.1167/jov.20.12.10

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: a review. *Neural Netw*. 113, 54–71. doi: 10.1016/j.neunet.2019.01.012

Phillips, W. A. (2015). Cognitive functions of intracellular mechanisms for contextual amplification. *Brain Cogn*. 112, 39–53. doi: 10.1016/j.bandc.2015.09.005

Phillips, W. A., Clark, A., and Silverstein, S. M. (2015). On the functions, mechanisms, and malfunctions of intracortical contextual modulation. *Neurosci. Biobehav. Rev*. 52, 1–20. doi: 10.1016/j.neubiorev.2015.02.010

Poirazi, P., Brannon, T., and Mel, B. W. (2003). Pyramidal neuron as two-layer neural network. *Neuron* 37, 989–999. doi: 10.1016/S0896-6273(03)00149-1

Poirazi, P., and Papoutsi, A. (2020). Illuminating dendritic function with computational models. *Nat. Rev. Neurosci*. 21, 303–321. doi: 10.1038/s41583-020-0301-7

Purushwalkam, S., Nickel, M., Gupta, A., and Ranzato, M. (2019). "Task-driven modular networks for zero-shot compositional learning," in *Proceedings of the IEEE International Conference on Computer Vision* (Long Beach, CA). doi: 10.1109/ICCV.2019.00369

Ramaswamy, S., and Markram, H. (2015). Anatomy and physiology of the thick-tufted layer 5 pyramidal neuron. *Front. Cell. Neurosci*. 9:233. doi: 10.3389/fncel.2015.00233

Rosch, E. (1975). Cognitive representations of semantic categories. *J. Exp. Psychol*. 104, 192–233. doi: 10.1037/0096-3445.104.3.192

Rosenbaum, C., Klinger, T., and Riemer, M. (2018). "Routing networks: adaptive selection of non-linear functions for multi-task learning," in *Proceedings of the 6th International Conference on Learning Representations* (Vancouver, BC).

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev*. 65:386. doi: 10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Rusu, A. A., Colmenarejo, S. G., Gülçehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., et al. (2016). "Policy distillation," in *Proceedings of the 4th International Conference on Learning Representations* (San Juan).

Sahni, H., Kumar, S., Tejani, F., and Isbell, C. (2017). "Learning to compose skills," in *Advances in Neural Information Processing Systems* (Long Beach, CA).

Schmidt-Hieber, C., Toleikyte, G., Aitchison, L., Roth, A., Clark, B. A., Branco, T., et al. (2017). Active dendritic integration as a mechanism for robust and precise grid cell firing. *Nat. Neurosci*. 20, 1114–1121. doi: 10.1038/nn.4582

Schoenfeld, G., Kollmorgen, S., Lewis, C., Bethge, P., Reuss, A. M., Aguzzi, A., et al. (2022). Dendritic integration of sensory and reward information facilitates learning. *bioRxiv*. doi: 10.1101/2021.12.28.474360

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv:1707.06347*. doi: 10.48550/arXiv.1707.06347

Sener, O., and Koltun, V. (2018). "Multi-task learning as multi-objective optimization," in *Advances in Neural Information Processing Systems* (Montréal, QC).

Sezener, E., Grabska-Barwínska, A., Kostadinov, D., Beau, M., Krishnagopal, S., Budden,D., et al. (2021). A rapid and efficient learning rule for biological neural circuits. *bioRxiv*. doi: 10.1101/2021.03.10.434756

Siegel, M., Körding, K. P., and König, P. (2000). Integrating top-down and bottom-up sensory processing by somato-dendritic interactions. *J. Comput. Neurosci*. 8, 161–173. doi: 10.1023/A:1008973215925

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362, 1140–1144. doi: 10.1126/science.aar6404

Snell, J., Swersky, K., and Zemel, R. S. (2017). "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems* (Long Beanch, CA).

Spruston, N. (2008). Pyramidal neurons: dendritic structure and synaptic integration. *Nat. Rev. Neurosci*. 9, 206–221. doi: 10.1038/nrn2286

Stuart, G., Spruston, N., and Häusser, M., editors (2016). *Dendrites, 3rd Edn*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198745273.001.0001

Stuart, G. J., and Spruston, N. (2015). Dendritic integration: 60 years of progress. *Nat. Neurosci*. 18, 1713–1721. doi: 10.1038/nn.4157

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction, 2nd Edn*. Cambridge, MA: MIT Press.

Takahashi, N., Ebner, C., Sigl-Glöckner, J., Moberg, S., Nierwetberg, S., and Larkum, M. E. (2020). Active dendritic currents gate descending cortical outputs in perception. *Nat. Neurosci*. 23, 1–9. doi: 10.1038/s41593-020-0677-8

van de Ven, G. M., and Tolias, A. S. (2019). Three scenarios for continual learning. *arXiv:1904.07734*. doi: 10.48550/arXiv.1904.07734

Veness, J., Lattimore, T., Budden, D., Bhoopchand, A., Mattern, C., Grabska-Barwinska, A., et al. (2021). "Gated linear networks," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence, Digital*.

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., and Behrens, T. E. J. (2022). How to build a cognitive map: insights frommodels of the hippocampal formation. *arXiv [Preprint]. arXiv: 2202.01682*. Available online at: https://arxiv.org/pdf/2202.01682.pdf

Wilson, A., Fern, A., Ray, S., and Tadepalli, P. (2007). "Multi-task reinforcement learning: a hierarchical Bayesian approach," in *Proceedings of the 24th International Conference on Machine Learning* (Corvallis, OR). doi: 10.1145/1273496.1273624

Wortsman, M., Ramanujan, V., Liu, R., Kembhavi, A., Rastegari, M., Yosinski, J., et al. (2020). "Supermasks in superposition," in *Advances in Neural Information Processing Systems, Digital*.

y Cajal, S. R. (1894). Neue darstellung vom histologischen bau des centralnervensystems. *Am. J. Psychol*. 6:450. doi: 10.2307/1411662

Yang, G., Lai, C. S. W., Cichon, J., Ma, L., Li, W., and Gan, W.-B. (2014). Sleep promotes branch-specific formation of dendritic spines after learning. *Science* 344, 1173–1178. doi: 10.1126/science.1249098

Yang, R., Xu, H., Wu, Y., andWang, X. (2020). "Multi-task reinforcement learning with soft modularization," in *Advances in Neural Information Processing Systems*.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. (2020). "Gradient surgery for multi-task learning," in *Advances in Neural Information Processing Systems*.

Yu, T., Quillen, D., He, Z., Julian, R., Narayan, A., Shively, H., et al. (2019). "Meta-world: a benchmark and evaluation for multi-task and meta reinforcement learning," in *Proceedings of the 3rd Conference on Robot Learning* (Osaka).

Zenke, F., Poole, B., and Ganguli, S. (2017). "Continual learning through synaptic intelligence," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney, NSW).

Zhang, Y., and Yeung, D.-Y. (2014). A regularization approach to learning task relationships in multitask learning. *ACM Trans. Knowl. Discov. Data* 8, 1–31. doi: 10.1145/2538028

**Conflict of Interest:** SA, AI, KG, and LS are employed by Numenta Inc. Numenta has some patents relevant to the work. Numenta has stated that use of its intellectual property, including all the ideas contained in this work, is free for non-commercial research purposes. In addition, Numenta has released all pertinent source code as open source under a GPL V3 license (which includes a patent peace provision).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Iyer, Grewal, Velu, Souza, Forest and Ahmad. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Computer Science

PERSPECTIVE
published: 06 May 2022
doi: 10.3389/fcomp.2022.846440

# Social Neuro AI: Social Interaction as the "Dark Matter" of AI

*Samuele Bolotta*[1] *and Guillaume Dumas*[2]*

[1] Department of Computer Science and Operations Research, Université de Montréal, Montreal, QC, Canada, [2] Department of Psychiatry, CHU Sainte-Justine Research Center, Mila-Quebec Artificial Intelligence Institute, Université de Montréal, Montreal, QC, Canada

This article introduces a three-axis framework indicating how AI can be informed by biological examples of social learning mechanisms. We argue that the complex human cognitive architecture owes a large portion of its expressive power to its ability to engage in social and cultural learning. However, the field of AI has mostly embraced a solipsistic perspective on intelligence. We thus argue that social interactions not only are largely unexplored in this field but also are an essential element of advanced cognitive ability, and therefore constitute metaphorically the "dark matter" of AI. In the first section, we discuss how social learning plays a key role in the development of intelligence. We do so by discussing social and cultural learning theories and empirical findings from social neuroscience. Then, we discuss three lines of research that fall under the umbrella of Social NeuroAI and can contribute to developing socially intelligent embodied agents in complex environments. First, neuroscientific theories of cognitive architecture, such as the global workspace theory and the attention schema theory, can enhance biological plausibility and help us understand how we could bridge individual and social theories of intelligence. Second, intelligence occurs in time as opposed to over time, and this is naturally incorporated by dynamical systems. Third, embodiment has been demonstrated to provide more sophisticated array of communicative signals. To conclude, we discuss the example of active inference, which offers powerful insights for developing agents that possess biological realism, can self-organize in time, and are socially embodied.

**Keywords: social interaction, cognitive architecture, virtual agents, social learning, Neuro-AI, neurodynamics, self-organization, Alan Turing**

OPEN ACCESS

***Edited by:***
*Adam Safron,*
*Johns Hopkins Medicine,*
*United States*

***Reviewed by:***
*James Derek Lomas,*
*Delft University of Technology,*
*Netherlands*
*Emily S. Cross,*
*Macquarie University, Australia*

***Correspondence:**
*Guillaume Dumas*
*guillaume.dumas@centraliens.net*

***Specialty section:***
*This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science*

***Received:*** *31 December 2021*
***Accepted:*** *11 April 2022*
***Published:*** *06 May 2022*

***Citation:***
*Bolotta S and Dumas G (2022) Social Neuro AI: Social Interaction as the "Dark Matter" of AI. Front. Comput. Sci. 4:846440. doi: 10.3389/fcomp.2022.846440*

## 1. THE IMPORTANCE OF SOCIAL LEARNING

### 1.1. Social Learning Categories

Various approaches have been proposed in order to reach a human-like level of intelligence. For example, some argue that scaling foundational models (self-supervised pretrained deep network models), data and compute can lead to such kind of intelligence (Bommasani et al., 2021; Yuan et al., 2022). Others argue that attention, understood as a dynamical control of information flow (Mittal et al., 2020), is all we need. Transformers have proposed a general purpose architecture where inductive biases shaping the flow of information are learned from the data itself (Vaswani et al., 2017); this architecture can be applied to various domains ranging from sequence learning to visual processing and time-series forecasting. Others argue that by having a complex enough environment, any

reward should be enough to elicit some complex behavior and end up in intelligent behavior that subserves the maximization of such reward. This therefore discards the idea that specialized problem formulations are needed for each ability (Silver et al., 2021). Our proposal stems from the idea that human cognitive functions such as theory of mind (the capacity to understand other people by ascribing mental states to them) and explicit metacognition (the capacity to reflect on and justify our behavior to others) are not genetically programmed, but rather constructed during development through social interaction (Heyes, 2018). Since their birth, social animals use their conspecifics as vehicles for gathering information that can potentially help them respond efficiently to challenges in the environment, avoiding harm and maximizing rewards (Kendal et al., 2018). Learning adaptive information from others results in better regulation of task performance, especially by gaining fitness benefits and in avoiding some of the costs associated with asocial, trial-and-error learning, such as time loss and energy loss as well as exposure to predation (Clark and Dumas, 2016). Importantly, cultural inheritance permeates a broad array of behavioral domains, including migratory pathways, foraging techniques, nesting sites and mates (Whiten, 2021). The spread of such information across generations gives social learning a unique role in the evolution of culture and therefore makes it a crucial candidate to investigate the biological bases of human cognition (Gariépy et al., 2014). In the current paper, we do not focus extensively on the differences between social learning in humans and in other animals as the cognitive processes used in acquiring behavior seem to be very similar across a wide range of species (Heyes, 2012). What sets humans apart from other animals, however, is: a) social learning in humans is highly rewarded from early infancy (Nielsen et al., 2012), b) the nature of the inputs surrounding humans is way more complex than for other animals (Heyes, 2012). According to the ontogenetic adaptation hypothesis (Tomasello, 2020), human infant's unique social-cognitive skills are the result of shared intentionality (capacity to share attention and intention) and are adaptations for life in a cultural group—with individuals coordinating, communicating and learning from each other in several ways. Recent reviews have identified four main categories of social learning that differ in what is socially learnt and in the cognitive skills that are required (Hoppitt and Laland, 2008; Whiten, 2021) (**Figure 1**). These categories have been developed through the approach of behaviorism. While we acknowledge that there is more to social learning than mere behavior (the affective and cognitive dimensions are equally crucial Gruber et al., 2021), we keep it as the focus of this short article because it is an empirically solid starting point with clarified mechanisms. The purpose of this section, then, is to give an example of social learning mechanisms that are common across multiple species and can be understood as a natural form of Social Neuro-AI. Moreover, this section aims at demonstrating how social interactions are a key component of biological intelligence; we make the case that they might be of inspiration for the development of socially intelligent artificial agents that can cooperate efficiently with humans and with each other. In other words, although there are examples of social agents (chatbots, non-player characters in video games, social robots), we argue that social interactions still remain the "dark matter" of the field. These social behaviors often emerge from a Piagetian perspective on human intelligence. As argued by Kovač et al. (2021), mainstream Deep Reinforcement Learning research sees intelligence as the product of the individual agent's exploration of the world; it mainly focuses on sensorimotor development and problems involving interaction with inanimate objects rather than social interactions with animate agents. This approach can and has given rise to apparent social behaviors, but we argue that this is not the best approach, as it does not involve any focus on the genuine social mechanisms *per se* (Dumas et al., 2014a). Instead, it sees social behaviors as a collateral effect of the intelligence of a solitary thinker. For this reason, as Schilbach et al. (2013) argued a decade ago that social interactions were the "dark matter" of cognitive neuroscience, here we argue that social interactions can also be considered metaphorically as the "dark matter" of AI (Schilbach et al., 2013). Indeed, more than being a rather unexplored topic, social interactions can constitute a critical missing piece for the understanding and modeling of advanced cognitive abilities.

At the most elementary level, enhancement consists of an agent observing a model that focuses on particular objects or locations and consequently adopting the same focus (Thorpe, 1963; Heyes, 1994). For example, it was demonstrated that bees outside the nest land more often on flowers that they had seen preferred by other bees (Worden and Papaj, 2005). This skill requires social agents to perform basic associative learning in relation to other agents' observed actions; it is likely to be the most widespread form of social learning across the animal kingdom. A more complex form of social learning consists of observational conditioning, which exposes a social agent to a relationship between stimuli (Heyes, 1994); this exposure causes a change in the agent. For example, the observation of experienced demonstrators facilitated the opening of hickory nuts by red squirrels, relative to trial-and-error learning (Weigl and Hanson, 1980). This is therefore a mechanism through which agents learn the value of a stimulus from the interaction with other agents. Yet a more complex form of social learning consists of affordance learning, which allows a social agent to learn the operating characteristics of objects or environments by observing the behavior of other agents (Whiten, 2021). For example, pigeons that saw a demonstrator push a sliding screen for food made a higher proportion of pushes than observers in control conditions, thus exhibiting affordance learning (Klein and Zentall, 2003). In other words, the animals perceive the environment partly in terms of the action opportunities that it provides. Finally, at the most complex level, copying another individual can take the shape of pure imitation, where every detail is copied, or emulation, where only a few elements are copied (Byrne, 2002). For example, most chimpanzees mastered a new technique for obtaining food when they were under the influence of a trained expert, whereas none did so in a population lacking an expert (Whiten, 2005). As to what is required for imitation, there are debates in the literature ranging from the distinctions between program-level and production-level imitation (Byrne, 2002) to the necessity of pairing Theory of Mind (ToM) with behavioral imitation to obtain "true" imitation (Call et al., 2005).

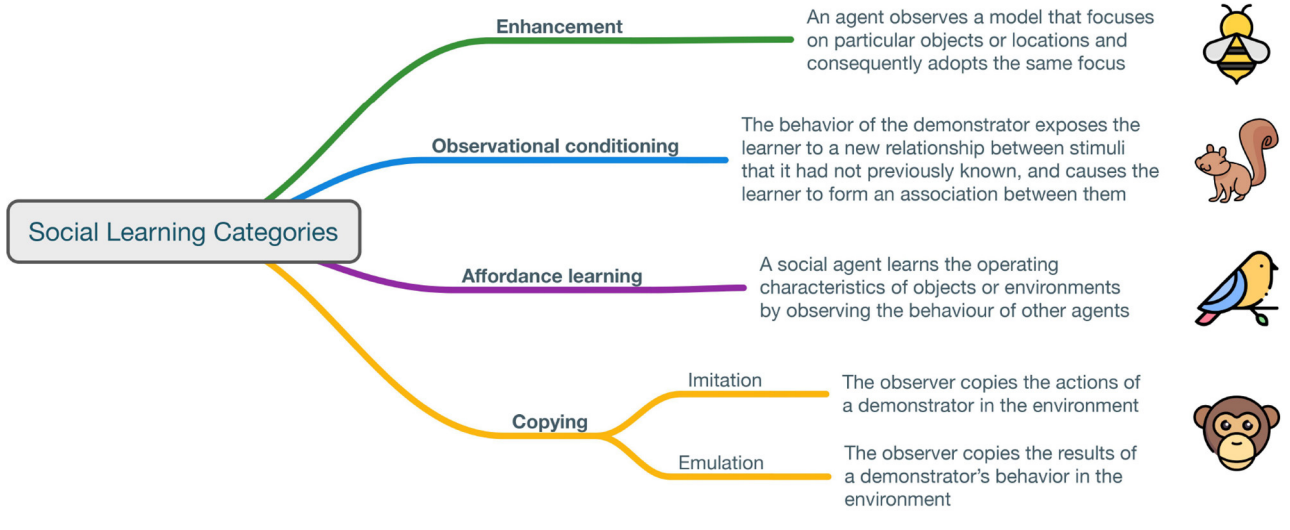



**FIGURE 1 |** Social learning categories. Figure inspired by Whiten (2005).

We refer the reader to Breazeal and Scassellati (2002) for a more detailed discussion of imitation in robots.

## 1.2. Social Learning Strategies

Crucially, while social learning is widespread, using it indiscriminately is rarely beneficial. This suggests that individuals should be selective in what, when, and from whom they learn socially, by following "social learning strategies" (SLSs; Kendal et al., 2018). Several SLSs might be used by the same population and even by the same individual. The aforementioned categories of social learning have been shown to be refined by modulating biases that can strengthen their adaptive power (Kendal et al., 2018). For example, an important SLS is copying when asocial learning would be costly; research has shown that, when task difficulty increases, various animals are more likely to use social information. Individuals also prefer using social information when they are uncertain about a task; high-fidelity copying is observed among children who lack relevant personal information (Wood et al., 2013). In general, other state-based SLSs can affect the decision to use social information, such as age, social rank, and reproductive state of the learner; for example, low- and mid-ranking chimpanzees are more likely to use social information than high-ranking individuals (Kendal et al., 2015). Model-based biases are another crucial category; for example, children prefer to copy prestigious individuals, where status is evidenced by their older age, popularity and social dominance (Flynn and Whiten, 2012). Multiple evidence also suggests that a conformist transmission bias exists, whereby the behavior of the majority of individuals is more likely to be adopted by others (Kendal et al., 2018).

### 1.2.1. Social Learning in Neuroscience

We have presented evidence that social learning is a crucial hallmark of many species and it manifests itself across different behavioral domains; without it, animals would lose the possibility to quickly acquire valuable information from their conspecifics and therefore lose fitness benefits. However, one important question is: how does the brain mediate social processes and behavior? Despite the progress made in social neuroscience and in developmental psychology, only in the last decade, serious efforts have started focusing on the answer to this question—as neural mechanisms of social interaction were seen as the "dark matter" of social neuroscience (Schilbach et al., 2013); recently, a framework for computational social neuroscience has been proposed, in an attempt to naturalize social interaction (Tognoli et al., 2018). At the intra-brain level, it was demonstrated that social interaction is categorically different from social perception and that the brain exhibits different activity patterns depending on the role of the subject and on the context in which the interaction is unfolding (Dumas et al., 2012). At the inter-brain level, functional Magnetic Resonance Imaging (fMRI) or Electroencephalography (EEG) recordings of multiple brains (i.e., hyperscanning) have allowed to demonstrate inter-brain synchronization during social interaction—specifically, while subjects were engaged in spontaneous imitation of hand movements (Dumas et al., 2010). Interestingly, the increase in coupling strength between brain signals was also shown to be present during a two-person turn-taking verbal exchange with no visual contact, in both a native or a foreign language context (Pérez et al., 2019). Inter-brain synchronization is also modulated by the type of task and by the familiarity between subjects (Djalovski et al., 2021). Overall, this shows that, beyond their individual cognition, humans are also coupled in the social dimension. Interestingly, the field of computational social neuroscience has also focused on explaining the functional meaning of such correlations between inter-brain synchronization and behavioral coupling. A biophysical model showed that the similarity of both endogenous dynamics and anatomical structure might facilitate inter-individual synchronization and explain our propensity to

socially bind with others via perception and actions (Dumas et al., 2012). More specifically, the connectome, a wiring diagram that maps all neural connections in the brain, not only facilitates the integration of information within brains, but also between brains. In those simulations, tools from dynamical systems thus suggest that beyond their individual cognition, humans are also dynamically coupled in the social realm (Dumas et al., 2012).

#### 1.2.2. Social Learning and Language Development

Regarding language development in humans, cognitive and structural accounts of language development have often conceptualized linguistic abilities as static and formal sets of knowledge structures, ignoring the contextual nature of language. However, good communication must be tailored to the characteristics of the listener and of the context—language can also be explained as a social construct (Whitehurst, 1978). For example, evidence shows that the language outcome of children with cochlear implants is heavily influenced by parental linguistic input during the first years after cochlear implantation (Holzinger et al., 2020). In terms of specific social learning variables, imitation has also been shown to play a major role in boosting language development, usually in the form of selective imitation (Whitehurst et al., 1974; Whitehurst and Vasta, 1975). Moreover, in children with autism spectrum disorder, social learning variables such as joint attention, immediate imitation, and deferred imitation have been shown to be the best predictors of language ability and rate of communication development (Toth et al., 2006). These results clearly suggest that social learning skills have an influence on language acquisition in humans.

## 2. STEPS TOWARD SOCIAL NEURO AI

How Could Social Learning Be Useful for AI?

In the previous sections, we have provided convincing evidence that interpersonal intelligence enhances intrapersonal intelligence through the mechanisms and biases of social learning. It is a crucial aspect of biological intelligence that possesses a broad array of modulating biases meant to strengthen its adaptive power. Recent efforts in computational social neuroscience have paved the way for a naturalization of social interactions.

Multi-agent reinforcement learning (MARL) is the best subfield of AI to investigate the interactions between multiple agents. Such interactions can be of three types: cooperative games (all agents working for the same goal), competitive games (all agents competing against each other), and mixed motive games (a mix of cooperative and competitive interactions). At each timestep t, each agent is attempting to maximize its own reward by learning a policy that optimizes the total expected discounted future reward. We refer the reader to high-quality reviews that have been written on MARL (Hernandez-Leal et al., 2019; Nguyen et al., 2020; Wong et al., 2021). Here, we highlight that, among others, low sample efficiency is one of the greatest challenges for MARL, as millions of interactions with the environment are usually needed for agents to learn. Moreover, multi-agent joint action space increases exponentially with the number of agents, leading to problems that are often intractable. In the last few years, part of the AI community has already started demonstrating that these problems can be alleviated by mechanisms that allow for social learning (Jaques, 2019; Ndousse et al., 2021). For example, rewarding agents for having a causal influence over other agents' actions leads to enhanced coordination and communication in challenging social dilemma environments (Jaques et al., 2019) and rewarding agents for coordinating attention with another agent improves their ability of coordination, by reducing the cost of exploration (Lee et al., 2021). More in general, concepts from complex systems such as self-organization, emergent behavior, swarm optimization and cellular systems suggest that collective intelligence could produce more robust and flexible solutions in AI, with higher sample efficiency and higher generalization (Ha and Tang, 2021). In the following sections, we argue that to exploit all benefits that social learning can offer AI and robotics, more focus on biological plausibility, social embodiment and temporal dynamics is needed. Studies have focused on the potential of conducting research at the intersection of some of these three axes (Kerzel et al., 2017; Husbands et al., 2021). Moreover, it is worth noticing that (Dumas et al., 2012; Heggli et al., 2019) offer a tentative glimpse of what the intersection of the three axes would look like-both using dynamical systems with computational simulations to address falsifiable scientific questions associated with the idea of social embodiment.

### 2.1. Biological Plausibility

Biological plausibility refers to the extent to which an artificial architecture takes inspiration from empirical results in neuroscience and psychology. The social learning skills and biases that we have shown so far are boosted in humans by their advanced cognitive architecture (Whiten, 2021). Equipping artificial agents with complex social learning abilities will therefore require more complex architectures that can handle a great variety of information efficiently. This is exactly what "Neuro-AI" aims at: drawing on how evolution has shaped the brain of humans and of other animals in order to create more robust agents (**Figure 2**). While the human unconscious brain aligns well with the current successful applications of deep learning, the conscious brain involves higher-order cognitive abilities that perform much more complex computations than what deep learning can currently do (Bengio, 2019). More specifically, "unconsciousness" is where most of our intelligence lies and involves unconscious abilities related to view-invariance, meaning extraction, control, decision-making and learning; "i-consciousness" is the part of human consciousness that is focused on integrating all available evidence to converge toward a single decision; "m-consciousness" is the part of human consciousness that is focused on reflexively representing oneself, utilizing error detection, meta-memory and reality monitoring (Graziano, 2017). Notably, recent efforts in the deep learning community have indeed focused on Neuro-AI: building advanced cognitive architectures that are inspired from neuroscience. In particular, the global workspace theory (GWT) is the most widely accepted theory of consciousness, and it postulates that when a piece of information is selected by attention, it may non-linearly

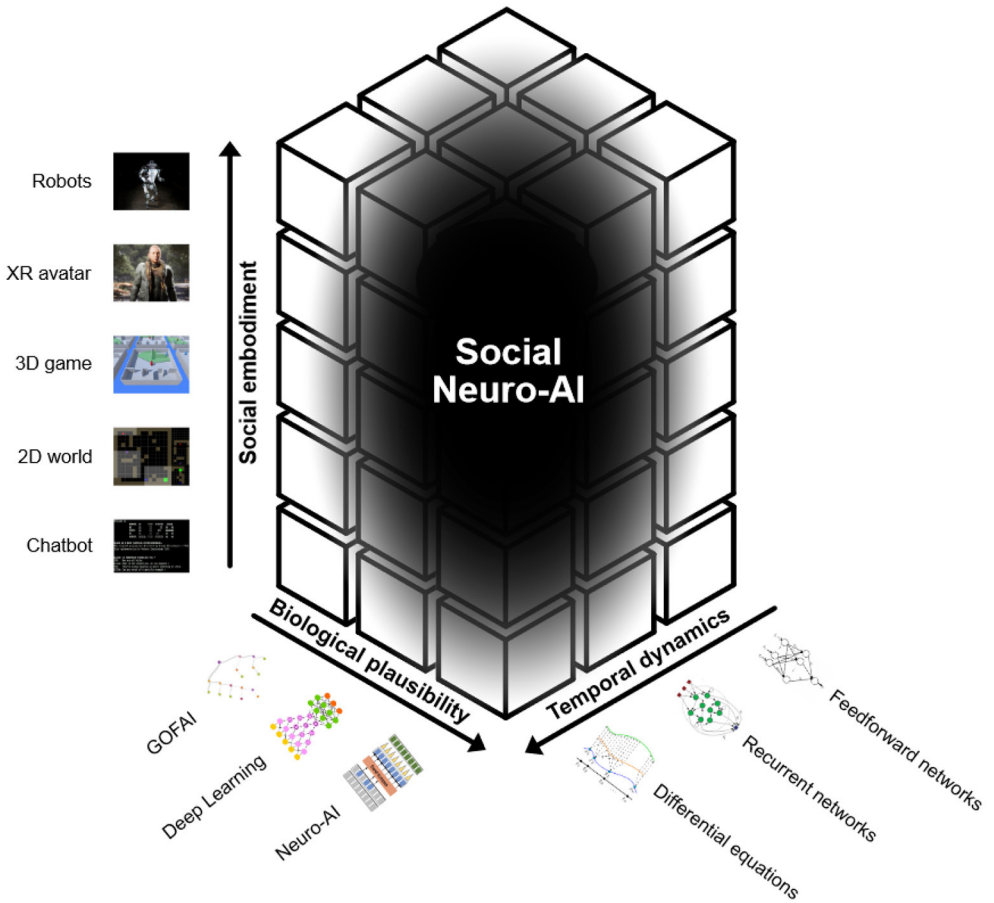



**FIGURE 2 |** Billions of humans interact daily with algorithms—yet AI is far from human social cognition. We argue that creating such socially aware agents may require "Social Neuro-AI"—a program developing 3 research axes: 1. Biological plausibility 2. Temporal dynamics 3. Social embodiment. Overall, those steps toward socially aware agents will ultimately help in aligned interactions between natural and artificial intelligence. Figure inspired by Schilbach et al. (2013).

achieve "ignition," enter the global workspace (GLW) and be shared across specialized cortical modules, therefore becoming conscious (Baars, 1993; Dehaene et al., 1998). The use of such a communication channel in the context of deep learning was explored for modeling the structure of complex environments. This architecture was demonstrated to encourage specialization and compositionality and to facilitate the synchronization of otherwise independent specialists (Goyal et al., 2021). Moreover, inductive biases inspired by higher-order cognitive functions in humans have been shown to improve OOD generalization. Overall, this section proposes that we draw inspiration from one structure we know is capable of comprehensive intelligence capable of perception, planning, and decision making: the human brain (**Figure 2**). For a more extensive discussion on biological plausibility in AI, we refer the reader to Hassabis et al. (2017) and Macpherson et al. (2021).

## 2.2. Temporal Dynamics

**Figure 2** more specifically, FFNs allow signals to travel only from input to output, whereas RNNs can have signals traveling in both directions and therefore introduce loops in the network. Incorporating differential equations in a RNN (continuous-time recurrent neural network) can help learn long-term dependencies (Chang et al., 2019) and model more complex phenomena, such as the effects of incoming inputs on a neuron. Moreover, viewing RNNs as a discretization of ordinary differential equations (ODEs) driven by input data has led to gains in reliability and robustness to data perturbations (Lim et al., 2021b). This becomes clear when one notices that many fundamental laws of physics and chemistry can be formulated as differential equations. In general, differential equations are expected to contribute to shifting the perspective from representation-centered to self-organizing agents (Brooks, 1991). The former view has been one predominant way of thinking about autonomous systems that exhibit intelligent behavior: such autonomous agents use their sensors to extract information about the world they operate in and use it to construct an internal model of the world and therefore rationally perform optimal decision making in pursuit of some goal. In other words, autonomous agents are information processing systems and their environment can be abstracted away as the source of answers to questions raised by the ongoing agents' needs.

Cognitive processes are thought to incorporate representational content and to acquire such contents via inferential processes instantiated by the brain. Importantly, according to this view, the sensorimotor connections of the agents to the environment are still relevant to understand their behavior, but there is no focus on what such connections involve and how they take place (Newell and Simon, 1976). The latter view, in line with the subsumption architecture introduced by Brooks (1991), shows how the representational approach ignores the nonlinear dynamical aspect of intelligence, that is, the temporal constraints that characterize the interactions between agent and environment. Instead, dynamics is a powerful framework that has been used to describe multiple natural phenomena as an interdependent set of coevolving quantitative variables (van Gelder, 1998) and a crucial aspect of intelligence is that it occurs in time and not over time. If we abstract away the richness of real time, then we also change the behavior of the agents (Smithers, 2018). In other words, one should indeed focus on the structural complexity and on the algorithmic computation the agents need to carry out, but without abstracting away the dynamical aspects of the agent-environment interactions: such dynamical aspects are pervasive and, therefore, necessary to explain the behavior of the system (van Gelder, 1998; Barandiaran, 2017; Smithers, 2018).

## 2.3. Social Embodiment

There has been a resurgence of enactivism in cognitive neuroscience over the past decade, emphasizing the circular causality induced by the notion that the environment is acting upon the individual and the individual is acting upon the environment. To understand how the brain works, then one has to acknowledge that it is embodied (Clark, 2013; Hohwy, 2013). Evidence for this shows that embodied intelligence in human children arises from the interaction of the child with the environment through a sensory body that is capable of recognizing the statistical properties of such interaction (Smith and Gasser, 2005). Moreover, higher primates interpret each other as psychological subjects based on their bodily presence; social embodiment is the idea that the embodiment of a socially interactive agent plays a significant role in social interactions. It refers to "states of the body, such as postures, arm movements, and facial expressions, that arise during social interaction and play central roles in social information processing." Thompson and Varela (2001) and Barsalou et al. (2003). This includes internal and external structures, sensors, and motors that allow them to interact actively with the world. We argue that robots are more socially embodied than digital avatars for a simple reason: they have a higher potential to use parts of their bodies to communicate and to coordinate with other agents (**Figure 2**). At a high level, sensorimotor capabilities in the avatar and robots are meant to model their role in biological beings: the agent now has limitations in the ways they can sense, manipulate, and navigate its environments. Importantly, these limitations are closely tied to the agent's function (Deng et al., 2019). The idea of social embodiment in artificial agents is supported by evidence of improvements in the interactions between embodied agents and humans (Zhang et al., 2016). Studies have shown positive effects of physical embodiment on the feeling of an agent's social presence, the evaluation of the agent, the assessment of public evaluation of the agent, and the evaluation of the interaction with the agent (Kose-Bagci et al., 2009; Gupta et al., 2021). In robots, social presence is a key component in the success of social interactions and it can be defined as the combination of seven abilities that enhance a robot/s social skills: 1. Express emotion, 2. Communicate with high-level dialogue, 3. Learn/recognize models of other agents, 4. Establish/maintain social relationships, 5. Use natural cues, 6. Exhibit distinctive personality and character, and 7. Learn/develop social competencies (Lee, 2006). Social embodiment thus equips artificial agents with a more articulated and richer repertoire of expressions, ameliorating the interactions with it (Jaques, 2019). For instance, in human-robot interaction, a gripper is not limited to its role in the manipulation of objects. Rather, it opens a broad array of movements that can enhance the communicative skills of the robot and, consequently, the quality of its possible interactions (Deng et al., 2019). The embodied agent is therefore the best model of the aspects of the world relevant to its surviving and thriving, through performing situationally appropriate actions (Ramstead et al., 2020) (**Figure 2**). Therefore, it will be crucial to scale up the realism of what the agents perceive in their social context, going from simple environments like GridWorld to more complex ones powered by video-game engines and, finally, to extremely realistic environments, like the one offered by the MetaHuman Creator of Unreal Engine. In parallel, greater focus is needed on the mental processes supporting our interactions with social machines, so as to develop a more nuanced understanding of what is 'social' about social cognition (Cross and Ramsey, 2021) and to gather insights critical for optimizing social encounters between humans and robots (Henschel et al., 2020). For a more extensive discussion on embodied intelligence, we refer the reader to Roy et al. (2021). These advancements will hopefully result in more socially intelligent agents and therefore in more fruitful interactions between humans and virtual agents.

# 3. ACTIVE INFERENCE

The active inference framework represents a biologically realistic way of moving away from rule-governed manipulation of internal representations to action-oriented and situationally appropriate cognition (Friston et al., 2006). More specifically, active inference can be seen as a self-organizing process of action policy selection (Ramstead et al., 2020), which a) concerns the selective sampling of the world by an embodied agent and b) instantiates in a generative model the goal of minimizing their surprise through perception and action (Ramstead et al., 2020). In other words, generative models do not encode exploitable and symbolic structural information about the world, because cognition does not perform manipulation of internal representations, but rather instantiates control systems that are expressed in embodied activity and utilize information encoded in the approximate posterior belief (Ramstead et al., 2020). Interestingly, by grounding GWT within the embodied perspective of the active inference framework, the Integrated

World Modeling Theory (IWMT) suggests that conscious experience can only result from autonomous embodied agents with global workspaces that generate integrative models of the world with spatial, temporal and causal coherence (Safron, 2020).

Active inference models are still very discrete in their architectures, especially regarding high-level cognitive aspects, but they may be a good class of models to raise the tension between computation and implementation (**Figure 2**). Therefore, they only have been able to handle small policies and state-spaces, while also requiring the environmental dynamics to be well known. However, using deep neural networks to approximate key densities, the agent can scale to more complex tasks and obtain performance comparable to common reinforcement learning baselines (Millidge, 2020). Moreover, one advantage of active inference is that the associated biologically inspired architectures predict future trajectories of the agent N steps forward in time, rather than just at the next step. By sampling from these trajectories, the variance of the decision is reduced (Millidge, 2020).

Interestingly, by grounding GWT within the embodied perspective of the active inference framework, the Integrated World Modeling Theory (IWMT) suggests that complexes of integrated information and global workspaces can entail conscious experiences if (and only if) they are capable of generating integrative world models with spatial, temporal, and causal coherence. These ways of categorizing experience are increasingly recognized as constituting essential "core knowledge" at the foundation of cognitive development (Spelke and Kinzler, 2007). In addition to space, time, and cause, IWMT adds embodied autonomous selfhood as a precondition for integrated world modeling.

## 4. A DETAILED PROPOSAL: HOW CAN INCREASED BIOLOGICAL PLAUSIBILITY ENHANCE SOCIAL AFFORDANCE LEARNING IN ARTIFICIAL AGENTS?

Attention has become a common ingredient in deep learning architectures. It can be understood as a dynamical control of information flow (Mittal et al., 2020). In the last decade, transformers have demonstrated how *attention may be all we need*, obtaining excellent performances in sequence learning (Vaswani et al., 2017), visual processing (Dosovitskiy et al., 2020) and time-series forecasting (Lim et al., 2021a). While transformers proposed a general purpose architecture where inductive biases shaping the flow of information are learned from the data itself, we can imagine a higher-order informational filter built on top of attention: an *Attention Schema (AS)*, namely a descriptive and predictive model of attention. In this regard, the attention schema theory (AST) is a neuroscientific theory that postulates that the human brain, and possibly the brain of other animals, does construct a model of attention: an attention schema (Graziano and Webb, 2015). Specifically, the proposal is that the brain constructs not only a model of the physical body but also a coherent, rich, and descriptive model of attention. The body schema contains layers of valuable information that help control and predict stable and dynamic properties of the body; in a similar fashion, the attention schema helps control and predict attention. One cannot understand how the brain controls the body without understanding the body schema, and in a similar way one cannot understand how the brain controls its limited resources without understanding the attention schema (Graziano, 2017). The key reason a higher-order filter on top of attention seems a promising idea for deep learning comes from control engineering: a good controller contains a model of the item being controlled (Conant and Ross Ashby, 1970). More specifically, a descriptive and predictive model of attention could help the dynamical control of attention and therefore maximize the efficiency with which resources are strategically devoted to different elements of an ever-changing environment (Graziano, 2017). Indeed, the performance of an artificial agent in solving a simple sensorimotor task is greatly enhanced by an attention schema, but its performance is greatly reduced when the schema is not available (Wilterson and Graziano, 2021). Therefore, the study of consciousness in artificial intelligence is not a mere pursuit of metaphysical mystery; from an engineering perspective, without understanding subjective awareness, it might not be possible to build artificial agents that intelligently control and deploy their limited processing resources. It has also been argued that, without an attention schema, it might be impossible to build artificial agents that are socially intelligent. This idea stems from the evidence that points at an overlap of social cognition functions with awareness and attention functions in the right temporo-parietal junction of the human brain (Mitchell, 2008). It was then proposed that an attention schema might also be used for social cognition, giving rise to an overlap between modeling one's own attention and modeling others' attention. In other words, when we attribute to other people an awareness of their surroundings, we are constructing a simplified model of their attention—a schema of others' attention (Graziano and Kastner, 2011). Indeed, such a model would enhance the ability of the agent to predict social affordances in real time, which is a goal the field has been trying to achieve in different ways (Shu et al., 2016; Ardón et al., 2021). Without a model of others' attention, even if we had detailed information about them, we could not predict their behavior on a moment-by-moment basis. However, with a component that tracks how and where other agents are focusing their resources in the environment, the probabilities for many affordances in the environment become computable in real time (Graziano, 2019). Specifically, there are three predictions that are investigated in this proposal. The first prediction is that, without an attention schema, attention is still possible, but it suffers deficits in control and thus leads to worse performance. The second prediction is that an attention schema is useful for modeling the attention of other agents as well —as the machinery that computes information about other people's attention is the same machinery that computes information about our own attention (Graziano and Kastner, 2011). The third prediction is that an agent equipped with an attention schema is going to have better OOD generalization than a classic Proximal Policy Optimization agent (Schulman et al., 2017), especially in environments in which the ability to

intelligently control and deploy limited processing resources is necessary.

## 5. CONCLUSION

At the crossroads of robotics, computer science, psychology, and neuroscience, one of the main challenges for humans is to build autonomous agents capable of participating in cooperative social interactions. This is important not only because AI will play a crucial role in daily life well into the future, but also because, as demonstrated by results in social neuroscience and evolutionary psychology, intrapersonal intelligence is tightly connected with interpersonal intelligence, especially in humans (Dumas et al., 2014b). In this opinion article, we have proposed an approach that unifies three lines of research that, at the moment, are separated from each other; in particular, we have proposed three research directions that are expected to enhance efficient exchange of information between agents. Biological plausibility attempts to increase the robustness and OOD generalization of algorithms by drawing on knowledge about biological brains; temporal dynamics attempts to better capture long-term temporal dependencies; social embodiment proposes that states of the body that arise during social interaction play central roles in social information processing. Unifying these axes of research would contribute to creating agents that are able to cooperate efficiently in extremely complex and realistic environments (Dennis et al., 2021), while interacting with other embodied agents and with humans.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

GD designed the theoretical framework and initial outline. SB wrote the first draft. SB and GD wrote the final version. Both authors contributed to the article and approved the submitted version.

## FUNDING

GD is funded by the Institute for Data Valorization (IVADO; CF00137433), Montreal, and the Fonds de recherche du Québec (FRQ; 285289).

## REFERENCES

Ardón, P., Pairet, R., Lohan, K. S., Ramamoorthy, S., and Petrick, R. P. A. (2021). Building affordance relations for robotic agents-a review. *arXiv:2105.06706 [cs]*. arXiv: 2105.06706. doi: 10.24963/ijcai.2021/590

Baars, B. J. (1993). *A Cognitive Theory of Consciousness*. Cambridge; Cambridge University Press. Google-Books-ID: 7w6IYeJRqyoC.

Barandiaran, X. E. (2017). Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. *Topoi* 36, 409–430. doi: 10.1007/s11245-016-9365-4

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., and Ruppert, J. A. (2003). "Social embodiment," in *Psychology of Learning and Motivation-Advances in Research and Theory* (San Diego, CA), 43–92.

Bengio, Y. (2019). The consciousness prior. *arXiv:1709.08568 [cs, stat]*. arXiv: 1709.08568.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv [Preprint]*. arXiv: 2108.07258. Available online at: http://arxiv.org/abs/2108.07258

Breazeal, C., and Scassellati, B. (2002). Robots that imitate humans. *Trends Cogn. Sci.* 6, 481–487. doi: 10.1016/S1364-6613(02)02016-8

Brooks, R. A. (1991). Intelligence without representation. *Artif. Intell.* 47, 139–159. doi: 10.1016/0004-3702(91)90053-M

Byrne, R. W. (2002). *Imitation of Novel Complex Actions: What Does the Evidence From Animals Mean*? San Diego, CA: Academic Press.

Call, J., Carpenter, M., and Tomasello, M. (2005). Copying results and copying actions in the process of social learning: chimpanzees (pan troglodytes) and human children (homo sapiens). *Anim. Cogn.* 8, 151–163. doi: 10.1007/s10071-004-0237-8

Chang, B., Chen, M., Haber, E., and Chi, E. H. (2019). Antisymmetricrnn: a dynamical system view on recurrent neural networks. *arXiv:1902.09689 [cs, stat]*. arXiv: 1902.09689.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, I., and Dumas, G. (2016). The regulation of task performance: a trans-disciplinary review. *Front. Psychol.* 6, 1862. doi: 10.3389/fpsyg.2015.01862

Conant, R. C., and Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system †. *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Cross, E. S., and Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human-machine interactions. *Trends Cogn. Sci.* 25, 200–212. doi: 10.1016/j.tics.2020.11.009

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14529–14534. doi: 10.1073/pnas.95.24.14529

Deng, E., Mutlu, B., and Mataric, M. (2019). Embodiment in socially interactive robots. *Foundat. Trends Rob.* 7, 251–356. doi: 10.1561/2300000056

Dennis, M., Jaques, N., Vinitsky, E., Bayen, A., Russell, S., Critch, A., et al. (2021). Emergent complexity and zero-shot transfer via unsupervised environment design. *arXiv:2012.02096 [cs]*. arXiv: 2012.02096.

Djalovski, A., Dumas, G., Kinreich, S., and Feldman, R. (2021). Human attachments shape interbrain synchrony toward efficient performance of social goals. *Neuroimage* 226, 117600. doi: 10.1016/j.neuroimage.2020.117600

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929.

Dumas, G., Chavez, M., Nadel, J., and Martinerie, J. (2012). Anatomical connectivity influences both intra- and inter-brain synchronizations. *PLoS ONE* 7, e36414. doi: 10.1371/journal.pone.0036414

Dumas, G., de Guzman, G. C., Tognoli, E., and Kelso, J. S. (2014a). The human dynamic clamp as a paradigm for social interaction. *Proc. Natl. Acad. Sci. U.S.A.* 111, E3726–E3734. doi: 10.1073/pnas.1407486111

Dumas, G., Laroche, J., and Lehmann, A. (2014b). Your body, my body, our coupling moves our bodies. *Front. Hum. Neurosci.* 8, 1004. doi: 10.3389/fnhum.2014.01004

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., and Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE* 5, e12166. doi: 10.1371/journal.pone.0012166

Flynn, E., and Whiten, A. (2012). Experimental "microcultures" in young children: identifying biographic, cognitive, and social predictors of information transmission: Identifying predictors of information transmission. *Child Dev*. 83, 911–925. doi: 10.1111/j.1467-8624.2012.01747.x

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Gariépy, J.-F., Watson, K., Du, E., Xie, D., Erb, J., Amasino, D., et al. (2014). Social learning in humans and other animals. *Front. Neurosci*. 8., 58 doi: 10.3389/fnins.2014.00058

Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., et al. (2021). Coordination among neural modules through a shared global workspace. *arXiv:2103.01197 [cs, stat*]. arXiv: 2103.01197.

Graziano, M. (2019). Attributing awareness to others: the attention schema theory and its relationship to behavioural prediction. *J. Consciousness Stud*. 26, 17–37. Available online at: https://grazianolab.princeton.edu/publications/attributing-awareness-others-attention-schema-theory-and-its-relationship

Graziano, M. S. A. (2017). The attention schema theory: a foundation for engineering artificial consciousness. *Front. Rob. AI* 4, 60. doi: 10.3389/frobt.2017.00060

Graziano, M. S. A., and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: a novel hypothesis. *Cogn. Neurosci*. 2, 98–113. doi: 10.1080/17588928.2011.565121

Graziano, M. S. A., and Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Front. Psychol*. 6, 500. doi: 10.3389/fpsyg.2015.00500

Gruber, T., Bazhydai, M., Sievers, C., Clément, F., and Dukes, D. (2021). The abc of social learning: affect, behavior, and cognition. *Psychol Rev*. doi: 10.1037/rev0000311. [Epub ahead of print].

Gupta, A., Savarese, S., Ganguli, S., and Fei-Fei, L. (2021). Embodied intelligence via learning and evolution. *Nat. Commun*. 12, 5721. doi: 10.1038/s41467-021-25874-z

Ha, D., and Tang, Y. (2021). Collective intelligence for deep learning: a survey of recent developments. *arXiv:2111.14377 [cs*]. arXiv: 2111.14377.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Heggli, O. A., Cabral, J., Konvalinka, I., Vuust, P., and Kringelbach, M. L. (2019). A kuramoto model of self-other integration across interpersonal synchronization strategies. *PLoS Comput. Biol*. 15, e1007422. doi: 10.1371/journal.pcbi.1007422

Henschel, A., Hortensius, R., and Cross, E. S. (2020). Social cognition in the age of human-robot interaction. *Trends Neurosci*. 43, 373–384. doi: 10.1016/j.tins.2020.03.013

Hernandez-Leal, P., Kartal, B., and Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Auton. Agents Multi Agent Syst*. 33, 750–797. doi: 10.1007/s10458-019-09421-1

Heyes, C. (2012). What's social about social learning? *J. Compar. Psychol*. 126, 193–202. doi: 10.1037/a0025180

Heyes, C. (2018). *Cognitive Gadgets*. Cambridge, MA: Harvard University Press.

Heyes, C. M. (1994). Social learning in animals: categories and mechanisms. *Biol. Rev. Camb. Philos. Soc*. 69, 207–231. doi: 10.1111/j.1469-185X.1994.tb01506.x

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press. Google-Books-ID: z7gVDAAAQBAJ.

Holzinger, D., Dall, M., Sanduvete-Chaves, S., Saldaña, D., Chacón-Moscoso, S., and Fellinger, J. (2020). The impact of family environment on language development of children with cochlear implants: a systematic review and meta-analysis. *Ear. Hear*. 41, 1077–1091. doi: 10.1097/AUD.0000000000000852

Hoppitt, W., and Laland, K. N. (2008). Chapter 3 social processes influencing learning in animals: a review of the evidence. *Adv. Study Behav*. 38, 105–165. doi: 10.1016/S0065-3454(08)00003-X

Husbands, P., Shim, Y., Garvie, M., Dewar, A., Domcsek, N., Graham, P., et al. (2021). Recent advances in evolutionary and bio-inspired adaptive robotics: exploiting embodied dynamics. *Appl. Intell*. 51, 6467–6496. doi: 10.1007/s10489-021-02275-9

Jaques, N. (2019). *Social and Affective Machine Learning* (Ph.D. thesis). Massachusetts Institute of Technology, Cambridge, MA, United States. Available online at: https://dspace.mit.edu/handle/1721.1/129901

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D. J., et al. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *arXiv:1810.08647 [cs, stat*]. arXiv: 1810.08647.

Kendal, R., Hopper, L. M., Whiten, A., Brosnan, S. F., Lambeth, S. P., Schapiro, S. J., et al. (2015). Chimpanzees copy dominant and knowledgeable individuals: implications for cultural diversity. *Evolut. Hum. Behav*. 36, 65–72. doi: 10.1016/j.evolhumbehav.2014.09.002

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., and Jones, P. L. (2018). Social learning strategies: bridge-building between fields. *Trends Cogn. Sci*. 22, 651–665. doi: 10.1016/j.tics.2018.04.003

Kerzel, M., Strahl, E., Magg, S., Navarro-Guerrero, N., Heinrich, S., and Wermter, S. (2017). "Nico-neuro-inspired companion: a developmental humanoid robot platform for multimodal interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Lisbon: IEEE), 113–120.

Klein, E. D., and Zentall, T. R. (2003). Imitation and affordance learning by pigeons (columba livia). *J. Comp. Psychol*. 117, 414–419. doi: 10.1037/0735-7036.117.4.414

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., and Nehaniv, C. L. (2009). Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Adv. Rob*. 23, 1951–1996. doi: 10.1163/016918609X12518783330360

Kovač, G., Portelas, R., Hofmann, K., and Oudeyer, P.-Y. (2021). Socialai: Benchmarking socio-cognitive abilities in deep reinforcement learning agents. *arXiv:2107.00956 [cs*]. arXiv: 2107.00956.

Lee, D., Jaques, N., Kew, C., Wu, J., Eck, D., Schuurmans, D., et al. (2021). Joint attention for multi-agent coordination and social learning. *arXiv:2104.07750 [cs*]. arXiv: 2104.07750.

Lee, K. (2006). Are physically embodied social agents better than disembodied social agents?: the effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *Int. J. Hum. Comput. Stud*. 64, 962–973. doi: 10.1016/j.ijhcs.2006.05.002

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021a). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast*. 37, 1748–1764. doi: 10.1016/j.ijforecast.2021.03.012

Lim, S. H., Erichson, N. B., Hodgkinson, L., and Mahoney, M. W. (2021b). Noisy recurrent neural networks. *arXiv:2102.04877 [cs, math, stat*]. arXiv: 2102.04877.

Macpherson, T., Churchland, A., Sejnowski, T., DiCarlo, J., Kamitani, Y., Takahashi, H., et al. (2021). Natural and artificial intelligence: a brief introduction to the interplay between ai and neuroscience research. *Neural Netw*. 144, 603–613. doi: 10.1016/j.neunet.2021.09.018

Millidge, B. (2020). Deep active inference as variational policy gradients. *J. Math. Psychol*. 96, 102348. doi: 10.1016/j.jmp.2020.102348

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cereb. Cortex* 18, 262–271. doi: 10.1093/cercor/bhm051

Mittal, S., Lamb, A., Goyal, A., Voleti, V., Shanahan, M., Lajoie, G., et al. (2020). "Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules," in *International Conference on Machine Learning* (PMLR), 6972–6986..

Ndousse, K. K., Eck, D., Levine, S., and Jaques, N. (2021). "Emergent social learning via multi-agent reinforcement learning," in *Proceedings of the 38th International Conference on Machine Learning,* eds M. Meila and T. Zhang, Vol. 139 (PMLR), 7991–8004. Available online at: https://proceedings.mlr.press/v139/ndousse21a.html

Newell, A., and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM* 19, 113–126. doi: 10.1145/360018.360022

Nguyen, T. T., Nguyen, N. D., and Nahavandi, S. (2020). Deep reinforcement learning for multi-agent systems: a review of challenges, solutions and applications. *IEEE Trans. Cybern*. 50, 3826–3839. doi: 10.1109/TCYB.2020.2977374

Nielsen, M., Subiaul, F., Galef, B., Zentall, T., and Whiten, A. (2012). Social learning in humans and nonhuman animals: theoretical and empirical dissections. *J. Comp. Psychol*. 126, 109–113. doi: 10.1037/a0027758

Pérez, A., Dumas, G., Karadag, M., and Duñabeitia, J. A. (2019). Differential brain-to-brain entrainment while speaking and listening in native and foreign languages. *Cortex* 111, 303–315. doi: 10.1016/j.cortex.2018.11.026

Ramstead, M. J., Kirchhoff, M. D., and Friston, K. J. (2020). A tale of two densities: active inference is enactive inference. *Adapt. Behav.* 28, 225–239. doi: 10.1177/1059712319862774

Roy, N., Posner, I., Barfoot, T., Beaudoin, P., Bengio, Y., Bohg, J., et al. (2021). From machine learning to robotics: challenges and opportunities for embodied intelligence. *arXiv:2110.15245 [cs]*. arXiv: 2110.15245.

Safron, A. (2020). An integrated world modeling theory (iwmt) of consciousness: Combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. *Front. Artif. Intell.* 3, 30. doi: 10.3389/frai.2020.00030

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., et al. (2013). Toward a second-person neuroscience. *Behav. Brain Sci.* 36, 393–414. doi: 10.1017/S0140525X12000660

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv:1707.06347 [cs]*. arXiv: 1707.06347.

Shu, T., Ryoo, M. S., and Zhu, S.-C. (2016). Learning social affordance for human-robot interaction. *arXiv:1604.03692 [cs]*. arXiv: 1604.03692.

Silver, D., Singh, S., Precup, D., and Sutton, R. S. (2021). Reward is enough. *Artif. Intell.* 299, 103535. doi: 10.1016/j.artint.2021.103535

Smith, L., and Gasser, M. (2005). The development of embodied cognition: six lessons from babies. *Artif. Life* 11, 13–29. doi: 10.1162/1064546053278973

Smithers, T. (2018). *Are Autonomous Agents Information Processing Systems*?, *1st Edn*. Hillsdale, NJ: Routledge.

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. *Develop. Sci.* 10, 89–96. doi: 10.1111/j.1467-7687.2007.00569.x

Thompson, E., and Varela, F. J. (2001). Radical embodiment: neural dynamics and consciousness. *Trends Cogn. Sci.* 5, 418–425. doi: 10.1016/S1364-6613(00)01750-2

Thorpe, W. H. (1963). *Learning and Instinct in Animals*. London: Methuen.

Tognoli, E., Dumas, G., and Kelso, J. A. S. (2018). A roadmap to computational social neuroscience. *Cogn. Neurodyn.* 12, 135–140. doi: 10.1007/s11571-017-9462-0

Tomasello, M. (2020). The adaptive origins of uniquely human sociality. *Philos. Trans. R. Soc. B Biol. Sci.* 375, 20190493. doi: 10.1098/rstb.2019.0493

Toth, K., Munson, J., N., Meltzoff, A., and Dawson, G. (2006). Early predictors of communication development in young children with autism spectrum disorder: joint attention, imitation, and toy play. *J. Autism. Dev. Disord.* 36, 993–1005. doi: 10.1007/s10803-006-0137-7

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* 21, 615–628. doi: 10.1017/S0140525X98001733

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Weigl, P. D., and Hanson, E. V. (1980). Observational learning and the feeding behavior of the red squirrel tamiasciurus hudsonicus: the ontogeny of optimization. *Ecology* 61, 213–218. doi: 10.2307/1935176

Whitehurst, G. J. (1978). The contributions of social learning to language acquisition. *Contemp. Educ. Psychol.* 3, 2–10. doi: 10.1016/0361-476X(78)90002-4

Whitehurst, G. J., Ironsmith, M., and Goldfein, M. (1974). Selective imitation of the passive construction through modeling. *J. Exp. Child Psychol.* 17, 288–302. doi: 10.1016/0022-0965(74)90073-3

Whitehurst, G. J., and Vasta, R. (1975). Is language acquired through imitation? *J. Psycholinguist. Res.* 4, 37–59. doi: 10.1007/BF01066989

Whiten, A. (2005). The second inheritance system of chimpanzees and humans. *Nature* 437, 52–55. doi: 10.1038/nature04023

Whiten, A. (2021). The burgeoning reach of animal culture. *Science* 372, eabe6514. doi: 10.1126/science.abe6514

Wilterson, A. I., and Graziano, M. S. A. (2021). The attention schema theory in a neural network agent: controlling visuospatial attention using a descriptive model of attention. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2102421118. doi: 10.1073/pnas.2102421118

Wong, A., Bäck, T., Kononova, A. V., and Plaat, A. (2021). Multiagent deep reinforcement learning: Challenges and directions towards human-like approaches. *arXiv:2106.15691 [cs]*. arXiv: 2106.15691.

Wood, L. A., Kendal, R. L., and Flynn, E. G. (2013). Copy me or copy you? the effect of prior experience on social learning. *Cognition* 127, 203–213. doi: 10.1016/j.cognition.2013.01.002

Worden, B. D., and Papaj, D. R. (2005). Flower choice copying in bumblebees. *Biol. Lett.* 1, 504–507. doi: 10.1098/rsbl.2005.0368

Yuan, S., Zhao, H., Zhao, S., Leng, J., Liang, Y., Wang, X., et al. (2022). A roadmap for big model. *arXiv [Preprint]*. arXiv: 2203.14101. Available online at: http://arxiv.org/abs/2203.14101

Zhang, M., Dumas, G., Kelso, J. S., and Tognoli, E. (2016). Enhanced emotional responses during social coordination with a virtual partner. *Int. J. Psychophysiol.* 104, 33–43. doi: 10.1016/j.ijpsycho.2016.04.001

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Bolotta and Dumas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fnbot.2022.846693

# Goal Oriented Behavior With a Habit-Based Adaptive Sensorimotor Map Network

*Felix M. G. Woolford* and Matthew D. Egbert*

*Artificial Life and Minds Lab, School of Computer Science, University of Auckland, Auckland, New Zealand*

We present a description of an ASM-network, a new habit-based robot controller model consisting of a network of adaptive sensorimotor maps. This model draws upon recent theoretical developments in enactive cognition concerning habit and agency at the sensorimotor level. It aims to provide a platform for experimental investigation into the relationship between networked organizations of habits and cognitive behavior. It does this by combining (1) a basic mechanism of generating continuous motor activity as a function of historical sensorimotor trajectories with (2) an evaluative mechanism which reinforces or weakens those historical trajectories as a function of their support of a higher-order structure of higher-order sensorimotor coordinations. After describing the model, we then present the results of applying this model in the context of a well-known minimal cognition task involving object discrimination. In our version of this experiment, an individual robot is able to learn the task through a combination of exploration through random movements and repetition of historic trajectories which support the structure of a pre-given network of sensorimotor coordinations. The experimental results illustrate how, utilizing enactive principles, a robot can display recognizable learning behavior without explicit representational mechanisms or extraneous fitness variables. Instead, our model's behavior adapts according to the internal requirements of the action-generating mechanism itself.

**Keywords: habit, sensorimotor contingencies, minimal cognition, robot controller, adaptive autonomy, enactivism**

OPEN ACCESS

***Edited by:***
*Inés Hipólito,*
*Humboldt University of Berlin, Germany*

***Reviewed by:***
*Ian Robertson,*
*University of Wollongong, Australia*
*Randall D. Beer,*
*Indiana University Bloomington, United States*

****Correspondence:***
*Felix M. G. Woolford*
*fmgwoolford@gmail.com*

***Received:*** *31 December 2021*
***Accepted:*** *28 March 2022*
***Published:*** *10 May 2022*

***Citation:***
*Woolford FMG and Egbert MD (2022) Goal Oriented Behavior With a Habit-Based Adaptive Sensorimotor Map Network. Front. Neurorobot. 16:846693. doi: 10.3389/fnbot.2022.846693*

## 1. INTRODUCTION

### 1.1. A Novel Habit-Based Controller

An enactive approach to AI and robotics requires us to take seriously the roots of autonomous agency and sense-making (Froese and Ziemke, 2009). To gain insight into the nature of intelligence, we cannot be content with mimicking the dynamics of intelligent behavior within the constraints of externally imposed norms. We must also ask *why* a system generates its own normative dimensions, how are they grounded in the material processes of the agent as a self-organizing system, and how do they relate to an intrinsically meaningful perspective on the world. These questions must motivate the design of our artificial models.

In recent years, a rich notion of habit as a core feature of cognition has been explored by theorists focussing on aspects of autonomy, sense-making, and anti-representationalism in enactivism (Barandiaran and Di Paolo, 2014; Egbert and Barandiaran, 2014; Barandiaran, 2017; Ramírez-Vizcaya and Froese, 2019; Hutto and Robertson, 2020). Of particular interest to us is a line of investigation concerning how habit serves as an approximation of a fundamental unit of the

sensorimotor domain of cognitive life, analogous to the role of the autopoetic cell as foundational to the biological domain of life (Buhrmann et al., 2013; Buhrmann and Di Paolo, 2017; Di Paolo et al., 2017; Di Paolo, 2019). In this view, a habit is a precarious but self-maintaining structure of sensorimotor activity, one that sustains itself as an entity over time by continually reproducing the conditions of its own performance.

Aspects of this view have been investigated through a computational model called the *Iterant Deformable Sensorimotor Medium* (IDSM) (Egbert and Barandiaran, 2014; Egbert and Cañamero, 2014; Egbert, 2018; Woolford and Egbert, 2019; Zarco and Egbert, 2019). The IDSM is essentially a mapping between a sensorimotor state and a change in motor state which is mutated as the medium is imprinted with a history of trajectories through a *sensorimotor space*. When coupled to a robot the medium serves as a controller which drives a kind of similarity-based behavior, in which the robot is driven to repeat the motor activity that it produced when it was historically in a similar sensorimotor state. As a behavior is repeated more frequently it in turn sustains and reinforces its influence on the IDSM mapping. Taken together, this facilitates the development of self-maintaining habitual behavior. Beyond the IDSM, a handful of other AI/robotics-type works drawing upon enactive theory have explored habits through different computational mechanisms or used comparable similarity-based mechanisms without being explicitly concerned with habit (Mirza et al., 2006; Iizuka and Di Paolo, 2007; Bedia et al., 2019; Georgeon and Riegler, 2019). Nevertheless, the scope of computational models of the enactive notion of habit remains relatively under-developed considering the relevance of habit to broader development of enactive cognitive science.

A recent criticism of the line of investigations working with the IDSM and related models is that they remain too minimal to provide an effective model of intelligent behavior, and that our artificial agents must be capable of developing an increasingly complex network of habits (Ramírez-Vizcaya and Froese, 2020). One of our recent works attempted to step in this direction by exploring how maintaining and refining a network of habits supported goal-oriented behavior acquired through evolutionary processes (Woolford and Egbert, 2020). Here we aim to push further in the direction of enriching the space of available computational models which can be used to explore habit-based cognition. To this end we present a new robot controller model, an *Adaptive Sensorimotor Map Network* (ASM-network). Building upon the kind of processes introduced with the IDSM, the ASM-network adaptively regulates the behavior of the robot as it engages with its environment, so as to maintain the viability of a structural organization within the model. That internal structure is motivated by the hypothesized organization of a sensorimotor agent as a structure of self-maintaining sensorimotor regularities (Di Paolo et al., 2017). The first half of this article thus details relevant elements of sensorimotor theory and adaptive sensorimotor agency, and then describes how the model captures some of these principles.

After presenting the model, we present an investigation to demonstrate its practical capacity as a tool for modeling cognition behavior. We investigate how a robot can solve a minimal cognition task previously investigated using evolutionary robotics methods (Beer, 1996). Evolutionary robotics methods have yielded invaluable developments in embodied theories of cognition through the analysis of the dynamics of adaptive behavior (Beer, 2008; Vargas et al., 2014). However, they have a critical limitation as an approach to investigating normativity and agency in an enactive sense, in that the viability constraints which the adaptive behavior maintains are externally imposed and have no meaningful correlation with the behavioral dynamics of the system. Barandiaran describes this as "the problem of dissociation between norm-establishing and norm-following processes" (Beer, 1997; Barandiaran and Egbert, 2014). Our investigation demonstrates that a system which attempts to reconcile these processes can still be used to investigate the same kinds of adaptive dynamics.

## 1.2. Sensorimotor Contingency Theory

Sensorimotor Contingency Theory is an attempt to account for the existence and quality of perceptual experience without appeals to notions of internal representation and other computational explanations (O'Regan and Noë, 2001; Noe, 2004). According to O'Regan and Noë's formulation of the theory, regularities in the relationship between movement and sensorimotor stimulation, and the "mastery" of such regularities, can explain an agent's phenomenal experience of perception as a result of their embodied activity. An archetypal example of the explanation provided by the theory is that of how the quality of "softness" is experienced. When a person squeezes a soft object such as a sponge, there is a particular contingent relationship between movements in the hand yielding a particular amount of pressure on the nerves in the finger tips. When squeezing a harder object such as a stone, the same muscle movements would coincide with a greater intensity of pressure on the fingertips. In mastering the laws of these relationships between motion and sensation the agent brings forth the experiences of softness and firmness, and the distinction between them. As O'Regan puts it, the experience of softness/firmness is a quality of the interaction in time between the body and the object, not an essential property of the object or "inside" the brain (O'Regan, 2011).

The theory's emphasis on perception as a process of active agent-environment interaction resonates with the enactive approach to cognition, especially with regard to the notion of sense-making. However the exact nature of sensorimotor contingencies and the notion of mastery in particular has proven challenging to reconcile with other aspects of enactivist thought. One challenge is the original formulation's apparent acceptance of cognitive representationalism to account for mastery (Hutto and Myin, 2012). Another is the question of how and why an autonomous agent would develop mastery of contingencies that are meaningful for that agent. Recently Di Paolo, Buhrmann, and Barandiaran provided a formalization of sensorimotor contingencies in terms of dynamical systems theory (Buhrmann et al., 2013). As part of this formalization, they defined four categories of sensorimotor contingencies, which describe different levels of the relationship between sensorimotor dynamics and the experience of the agent:

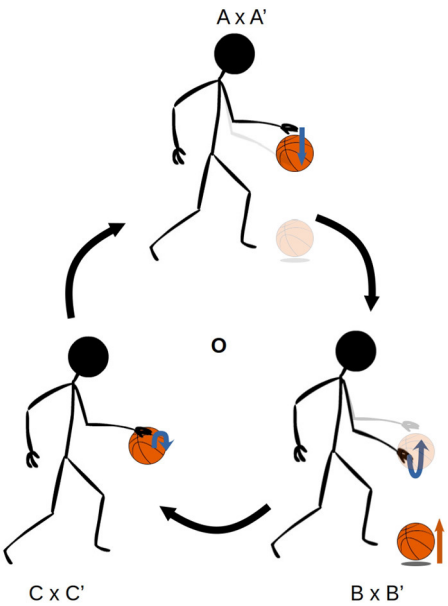



**FIGURE 1 |** The sensorimotor scheme **O**, associated with the behavior of bouncing a basketball. The scheme consists of the cyclical organization of the three coordinations, A × A' → B × B' → C × C' → A × A', where the arrows indicate the transitional structure between these coordinations. The A × A' notation refers to the simultaneous realization of the agent-side sensorimotor support structure A, and the environment-side response structure A'.

1. **Sensorimotor (SM) Environment**, the set of all regularities in the way that actions may affect perceptions for a given body in a given environment, e.g., between eye movements and retinal stimulations, without regard for the agent's internal dynamics involved in performing those actions.
2. **Sensorimotor Habitat** regular trajectories within the sensorimotor environment associated with a particular agent's way of being. In other words, the time-extended regularities involved in the loop of both action affecting perception *and* perception affecting action, given the specific internal properties of the acting agent.
3. **Sensorimotor Coordination**, a clustering of regularities within the sensorimotor habitat associated with the fulfillment of a particular goal of an agent.
4. **Sensorimotor Scheme**, an organization of coordinations associated with a particular normative framework and modulated according to that framework.

These categories clarify the distinction between (1) contingencies considered in more mechanical or statistical terms relating to the coupling between body and environment, and (2) contingencies as related to the experience of the agent in terms of its needs and expectations. We will briefly expand upon the details of these categories which are most relevant to this investigation. We are mostly concerned here with what it means for sensorimotor coordinations to be organized in relation to *goals* and *norms*.

**Figures 1**, **2** illustrate the way in which the bodily and environmental aspects of bouncing a basketball relate to the ideas of sensorimotor coordination and sensorimotor schemes. We can consider this as a **scheme** composed of three **coordinations**: Pushing the basketball toward the ground; preparing to receive the ball as it bounces on the ground; and receiving the ball as it returns to the hand. Each of these coordinations describes a particular class of embodied dynamics, all associated with a particular aspect of the basketball bouncing process. Assuming the scheme is stable, then each particular instance of enacting this scheme will follow the same sequence of coordinated acts, with each instance of a coordination varying in its precise dynamics but reliably establishing the enabling conditions for an instance of the next coordination. The processes involved in these transitions are honed over time with respect to various normative dimensions associated with bouncing a basketball effectively and efficiently.

A crucial emphasis of this formalization is that these regularities are not just concerned with the agent's brain and body, but involve the entirety of the brain-body-world system. The regularities associated with the performance of this scheme encompass both the positioning and readiness of the agent's body, and the position of the ball in relation to the body. **Figure 2** illustrates how these regularities form a sensorimotor-coordination. Each coordination encompasses co-occurring regularities in the dynamics of both the agent (i.e., the actions and sensations associated with pushing the ball downwards, in the case of A here) and the environment (i.e., the position of the ball in space and its physical attributes, in the case of A'), within a specific temporal context with respect to several other coordinations. In other words, every instance of a particular coordination is a trajectory through a space of sensorimotor states (sensorimotor space), over which relevant state variables are transformed from one particular set of enabling conditions to another set, and a coordination structure ultimately is composed of the infinite set of possible variations on these trajectories.

In this example, we have outlined what an organization of sensorimotor contingencies might look like with respect to a particular activity, but not how or why such an organization would develop. Exactly whose goals and norms are we referring to when we say a scheme is associated with a particular normative framework, and where do those norms come from?

## 1.3. Sensorimotor Agency

Di Paolo et al. (2017) integrated their formalization of SMCs with a proposal for an account of cognition in which a complex embodied agent, such as a human, is not reducible to just its biological processes, but rather consists of many autonomous processes in deeply interwoven but ultimately irreducible biological, behavioral, and social domains. These processes and the relations between them ultimately ground the goals and norms which are relevant to the higher-level categories of sensorimotor contingencies. The core of their proposal is the idea that an organization of sensorimotor contingencies can manifest the necessary and sufficient properties to possess its own form of agency. Such an organization is proposed to constitute behavioral domain's analog to the notion of the cellular organism as biological agent.

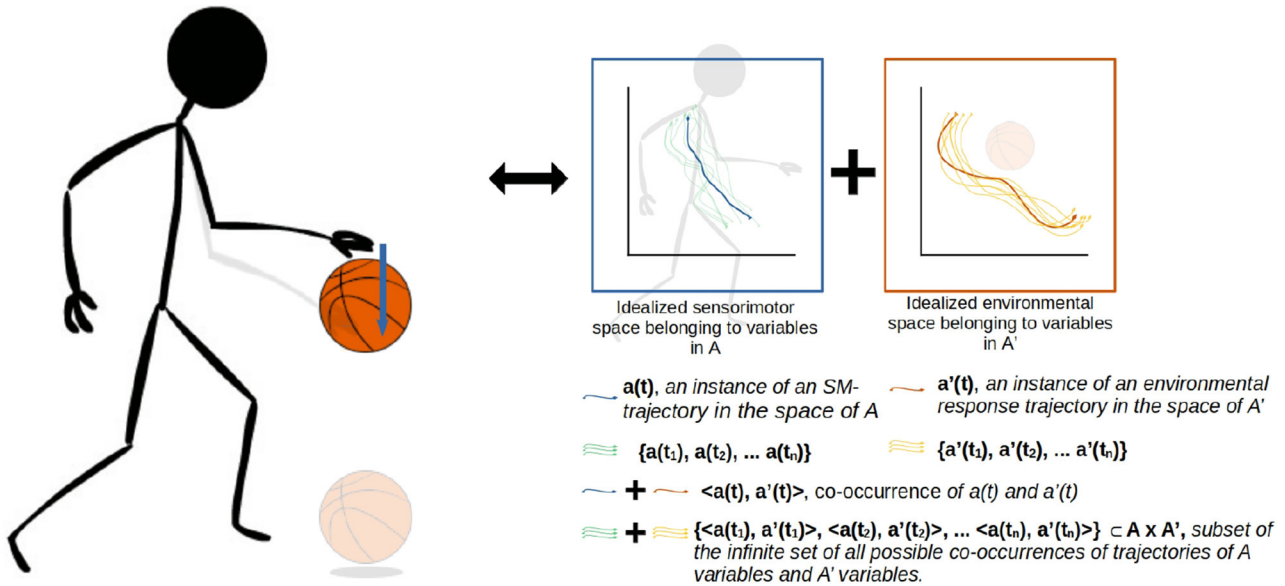



**FIGURE 2 |** A more detailed visualization of the A × A' coordination from the previous figure. The coordination captures not just the agent's actions (A) and sensations but also the environmental processes that happen concurrently (A'). The state spaces represent theoretical projections of the state spaces of the relevant variables on the agent-side and environment-side of the engagement. The trajectories are not just co-occurring but also circularly causal, with the sensorimotor trajectory being a function of both internal (e.g., neuromuscular) dynamics and the environmental impacts on the agent's body, and similarly the environmental trajectory being a function of both the agent's actions and environmental processes such as gravity acting upon the ball.

The short version of the definition of an agent that underlies the proposal is:

> An autonomous system capable of adaptively regulating its coupling with the environment according to the norms established by its own viability condition (Barandiaran et al., 2009).

In the case of a sensorimotor agent, this system is a self-individuating, self-sustaining organization of activity which emerges within the dynamics of a brain-body-environment system, an entity composed of interacting sensorimotor schemes. This interaction refers to the relations between sensorimotor schemes in time—the way in which the performance of one scheme can regularly support, inhibit, or require the performance of other activities. At a high level we can think of each of these schemes as the regularities concerning a particular embodied activity: drinking from a cup, walking, reaching for a phone. Crucially, these are regularities which emerge not just in the dynamics of the internal process of the agent, but over the entire coupled system comprising the physical properties of the world, the agent's body, and the agent's neurological and physiological dynamics. A structure of interrelated activities can be understood as constituting its own kind of entity in the sensorimotor domain. Such a entity would comprise the entirety of the activities involved in a particular embodied agent's mode of being. The self-individuation of this structure refers to the way in which the stability of this structure is established through the very processes of activity that constitute it. These processes establish an operational closure of all of those activities which stabilize support for other activities within this structure, and in turn depend on the support of other activities in the structure. This process of self-individuation grounds a dimension of normativity related to the continuation of the activities which constitute the sensorimotor agent, as well as to the integrity of the structural relationships between activities. Actions and environmental structures may take on meaning of being more or less good or bad depending on how they support or disrupt that process. These elements may be irrelevant or even in direct opposition to the agent's viability at another level, such as the biological. If the dynamics of the brain-body-environment coupling are such that the behavior of the agent may change and develop to maintain its sensorimotor organization according to these norms, then we have an autonomous structure at the sensorimotor level which adaptively regulates its engagement with its environment, thus fulfilling the criteria of an organization that possesses its own form of agency.

This theory of sensorimotor agency has the potential to explain how and why an agent develops the sensorimotor mastery necessary to ground its phenomenal experience of the world, and to explain how complex behaviors and skills can take on "a life of their own," apparently divorced from any role in maintaining the viability of the biological agent engaged in those behaviors. Clearly though, the idea of an agent constituted by its own acts presents a challenging conceptual puzzle (Di Paolo et al., 2017, Chapter 6). Artificial models have a key role to play in both clarifying and developing this and associated theories. Much of this work to date has focussed on the notion of

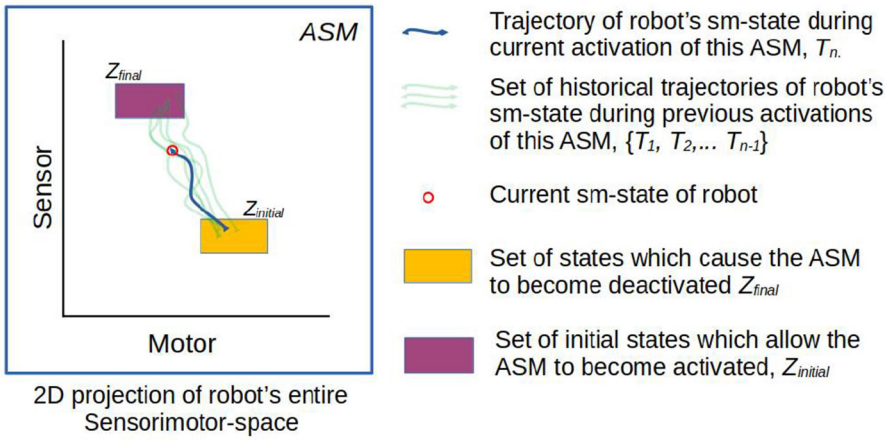



**FIGURE 3 |** Simplified visualization of an ASM-unit in operation. The model essentially stores a number of historical trajectories in this space which have passed from the within range initial state to the range of final states, and utilizes information about those trajectories to generate motor activity for the robot in its current state.

*habit*, which provides a useful "first approximation" (Egbert and Barandiaran, 2014) of a minimal kind of self-sustaining sensorimotor entity. A habit may be conceived of as a dissipative structure of activity which depends upon its own continual re-performance for stability. In the context of the formalization of sensorimotor contingency categories, the structure of a minimal habit is akin to a single, circular scheme in which a series of coordinations ultimately reproduce the conditions for their own re-enactment. This structure grounds a single normative dimension concerning that continuing cycle of reproducing enabling conditions. Although the notion of habit—especially a single habit in isolation—does not capture the full richness of sensorimotor agency (Di Paolo et al., 2017, p. 146–154), it provides a starting point for investigation.

This brings us to our own work. Our aim is to build upon previous models that have been used to investigate this kind of enactive notion of habit, moving a step closer to the idea of sensorimotor agency proper. In particular our model aims to investigate the notion of habit more directly in terms of those categories of sensorimotor contingency, by explicitly incorporating properties of sensorimotor structure and dynamics which support the maintenance of that structure's viability in the face of environmental disruptions and obstacles. We now present a description of this model.

## 2. MODEL

### 2.1. An Overview of the ASM-Network Model

In the simplest description, the ASM-network model is a robot controller which generates motor commands for a robot based on the relationship between its current sensorimotor state and its history of sensorimotor trajectories. It consists of a network of Adaptive Sensorimotor Map units (ASM-units). The general design of each unit is similar to an earlier model, the Iterant Deformable Sensorimotor Medium (IDSM) (Egbert and Barandiaran, 2014; Egbert and Cañamero, 2014), while the mechanisms involved in organizing these units as a network are based on our previous Sensorimotor Sequence Reiterator model (Woolford and Egbert, 2020). Both of those models, and this one, may be considered as belonging to a family of habit-based robot controllers. These models are similar in two primary ways: Firstly, they are all specifically concerned with a sensorimotor level of abstraction (i.e., leaving aside lower level neural and physiological dynamics). Secondly, when coupled to the motors and sensors of an embodied robot as a controller, they serve to encourage the repetition and reinforcement of the robot's historical behaviors. The ASM-network is unique among these controllers in that it monitors the way in which new performances affect the stability of historically established behaviors, and *adaptively* modulates its own dynamics in the direction of maintaining the viability of those behaviors. Additionally, the processes of the model are organized analogously to the organization of sensorimotor contingencies in an autonomous sensorimotor entity as we described in the previous section.

**Figures 3**, **4** illustrate the basic elements of the ASM-network model. In operation, only one ASM-unit is "active" at any one time, and that unit is responsible for governing the changes in motor activity of the robot. This state of activation traverses the network over time. As a rough approximation, we may think of an individual ASM-unit as being associated with the agent dynamics associated with a single sensorimotor coordination structure, and a collection of these coordinations in a network as being associated with a sensorimotor scheme. **Figure 5** illustrates a hypothetical relationship between our basketball-bouncing sensorimotor scheme and an instantiated ASM-network model in the context of a robot, controlled by an ASM-network, which is able to successfully enact that sensorimotor scheme. The model components illustrated there will become clear as we discuss further.

We now discuss the model in three parts: Firstly, we explain model at the level of individual ASM-units, and then

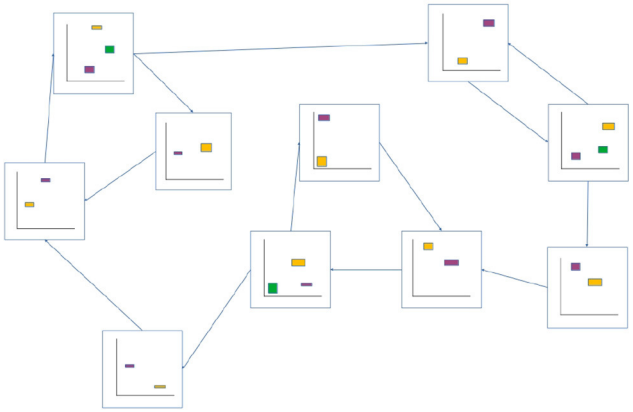


**FIGURE 4 |** Example of an ASM-network consisting of 10 ASM-units. Like in **Figure 3**, ASM-units are represented by two-dimensional projections of sensorimotor-spaces. Arrows indicate that activation will transition from one ASM-unit to another when the controlled robot is in an appropriate sensorimotor-state. Note that the yellow initial-state regions of each ASM-unit corresponds in space with a purple or green final-state region from a preceding ASM-unit.

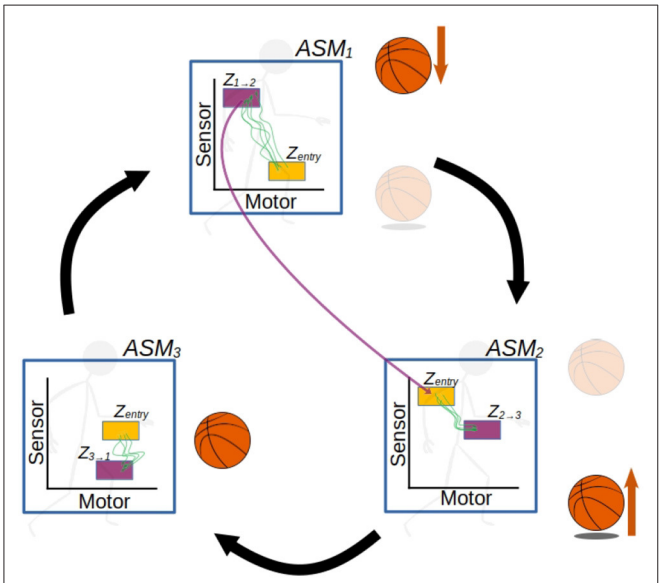



**FIGURE 5 |** The activity of an ASM-network with three ASM-units, overlaying a visualization of the robot engaged in the ball-bouncing scheme. Each ASM-unit simulates a particular sequential component of the dynamics involved in the robot's side of the coupling. Each activation of the ASM-unit yields the necessary state for the activation of the next ASM-unit in the sequence. The progression of activations through the network mirrors the temporal arrangement of coordinations in a scheme.

at the network level. Finally, we will explain how these two levels interact to adaptively maintain stable behavior. Symbols used in the following sections are summarized in **Table 1**.

**TABLE 1 |** Symbols for model parameters and components.

| Symbol | Value for section 3 | Description |
|---|---|---|
| $\tau$ | 0.1 s | Period between node-creation events in an ASM-unit |
| $N_{max}$ | 8,000 | Maximum number of nodes in a single ASM-unit before nodes begin to be replaced |
| $w$ | 1.5 | Scales the relative importance of $P$ and $V$ comparisons in similarity metric |
| $t_g$ | 8 s | Period of activation for an ASM-unit before non-historical transitions may occur |
| $t_h$ | 16 s | Maximum period of activation for an ASM-unit |
| $d_{cutoff}$ | 0.2 sm-space units | Maximum distance in SM-space for candidate parent nodes |
| $N$ | $\langle P, V, \Delta m, C, Z, A \rangle$ | ASM-unit node |
| $P$ | 8-dimensional vector | Position in SM-space of a node |
| $V$ | 8-dimensional vector | Displacement of node's position from previously created node's position |
| $\Delta m$ | 1-dimensional vector | Change-in-motor-state generated at node-creation |
| $C$ | label | Node class label, inherited from parent node |
| $Z$ | label | Label of the transition condition which terminates the activation of an ASM-unit |
| $A$ | 0 or 1 | Flag which marks a node as reinforced or inhibited |

*Described values for the parameters and dimensionality of vectors are those used in the investigation in Section 3 but are only indicative of a suitable order of magnitude for the general case.*

## 2.2. ASM Unit-Level Architecture

All ASM-units shares the same functional properties, and are essentially self-contained in their operation in most cases. Therefore, we can present most of the model in terms of a single ASM-unit in isolation from the rest of the network. **Figure 3** illustrates the basic elements of an ASM-unit graphically. Readers familiar with the IDSM will recognize several core similarities in the ASM-unit's architecture.

A key concept at the heart of an ASM-unit is the notion of the *sensorimotor space*, the construct of all possible values of *all* sensor and motor variables of the controlled robot, which are each treated as bounded scalars. Conceptually we may think of these values as representing the full range of movements and sensations accessible to the robot. At any moment, the robot's sensorimotor state is the value of all of those sensors and motors:

$$\mathbf{sm}(t) = \begin{bmatrix} m_1(t) \\ \vdots \\ m_n(t) \\ s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} \quad (1)$$

The ASM-unit essentially operates in terms of comparing the current sensorimotor state to historical states in terms of their

position in sensorimotor space. In an example in the context of the basketball-bouncing scheme, we may think of it comparing a particular movement and sensation of the arm to other historical movements and sensations.

The ASM-unit also gives primacy to the concept of the time-extended trajectory of the robot's sensorimotor state through sensorimotor space. In the context of each ASM-unit, which is only active for a finite segment of time, we specifically use the term *sensorimotor trajectory* to refer to discrete segments of the robot's trajectory through the space, beginning at the time of the ASM-unit's activation and ending at the time of its termination. The ASM-unit therefore has a collection of historical sensorimotor trajectories, based on how many times it has been activated. **Figure 3** illustrates a collection of five historical sensorimotor trajectories, suggesting that it the ASM-unit is currently in its sixth activation. As per the basketball example, we may think of each of these sensorimotor trajectories as instances of the robot's sensorimotor activity as it was going through a particular performance of a particular act, e.g., of pushing the ball.

An ASM-unit is designed so that it causes historical sensorimotor trajectories to be repeated, by dynamically generating a sensorimotor-state to change-in-motor-state map, $f(\mathbf{sm}) = \dot{m}$, based on those historical sensorimotor trajectories. The dynamics of the model are precarious in that information of historical trajectories is lost over time, so for a particular behavior to be sustained over the long term it must regularly recur. However, repeating historical trajectories is not as simple as merely repeating historical motor actions. The time-extended evolution of the sensorimotor state may be separated into the evolution of the motor states and evolution of the sensor states:

$$\dot{\mathbf{sm}} = \dot{m} + \dot{s} = \begin{bmatrix} \dot{m}_1 \\ \vdots \\ \dot{m}_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \dot{s}_1 \\ \vdots \\ \dot{s}_n \end{bmatrix} \tag{2}$$

$$\begin{aligned} \dot{m} &= f(\mathbf{sm}) \\ \dot{s} &= g(m, \mathbf{e}) \end{aligned} \tag{3}$$

Where, $\mathbf{e}$ is a vector representing the environmental state (i.e., properties of the world and the robot's position in it). Ultimately the ASM-unit is only responsible for generating $f(\mathbf{sm})$ (see later, Equations 5–7), but has no direct influence on $g(m, \mathbf{e})$. In other words, the ASM-unit is designed to reproduce historical sensorimotor trajectories, but it only has direct control over the change in state in a subset of the relevant dimensions. The same motor action in two different contexts may yield different sensorimotor trajectories depending on the environmental state. This produces a tension which causes only certain behaviors to be stable—those in which the repetition of certain sensorimotor states is concurrent with the repetition of certain environmental states. In our basketball example, this means that regular movements and sensations are only stable if the physical properties of the ball bouncing off the ground are also regular. This challenge relates to the concurrence of agent-side and environment-side dynamics in a sensorimotor coordination as illustrated in **Figure 2**.

In any non-trivial system, natural variations in the environmental state will mean that exact repetitions of historical trajectories are not possible, and thus the ASM-unit needs a mechanism for comparing the relative similarity of the current state to historical states. Thus, the influence of particular trajectories through sensorimotor space propagates over the entire state space, such that historical change-in-motor-state commands are adjusted for the current context. From a design perspective, the functional effectiveness of these comparisons and adjustments (i.e., the comparisons are accurate and the adjustments suitable) are critical to the ASM-unit's ability to repeat historical behaviors.

This brings us to the two operational mechanisms of an ASM-unit, (1) storing information about historical sensorimotor trajectories, and (2) using that information to generate change-in-motor-state commands. Sensorimotor trajectories are sampled at discrete intervals and stored by the ASM-unit as sequences of nodes, each representing the sensorimotor state of the robot at the moment of sampling. When determining a change-in-motor-state command, the ASM-unit compares the current SM-state to these stored nodes, and generates an output based on the state of the most similar stored node. The remainder of this subsection will explain the details of these two basic mechanisms.

### 2.2.1. Node Creation

The controller is applied in a simulation of a continuous-time system, using the Euler method to approximate continuous dynamics, with a time step of size 0.01. At regular intervals of $\tau$ time units ($\tau = 0.1s$) a node is created to store the current state of the robot, and a vector $\Delta$m is generated which determines the rate of change to the robot's motor state over the next $\tau$ interval. The structure of a node is illustrated in **Figure 6**. Upon creation, each node stores the following information:

1. The current sensorimotor state of the robot, which we regard as the node's position in sensorimotor space $P$,
2. The vector $V$ from the position of the previously created node to the position of the current node,
3. The vector $\Delta$m for the intended change in motor state determined during node creation. The process of generating $\Delta$m is discussed shortly.
4. An identifying class label for the node, $C$, which is inherited from the most similar historical node (parent node) and will propagate to future similar nodes. This too is discussed further shortly.

There are also data stored in each node that relate to how the ASM-unit is exited. When the activation of an ASM-unit ends, all of those nodes which were created during that activation are modified to include:

5. An identifying label, $Z$, of the exit region that caused the activation to transition to another ASM-unit.

Finally, when the activation of the *next* ASM-unit is completed, every created node is updated with feedback regarding the controller's progression through the higher order network:

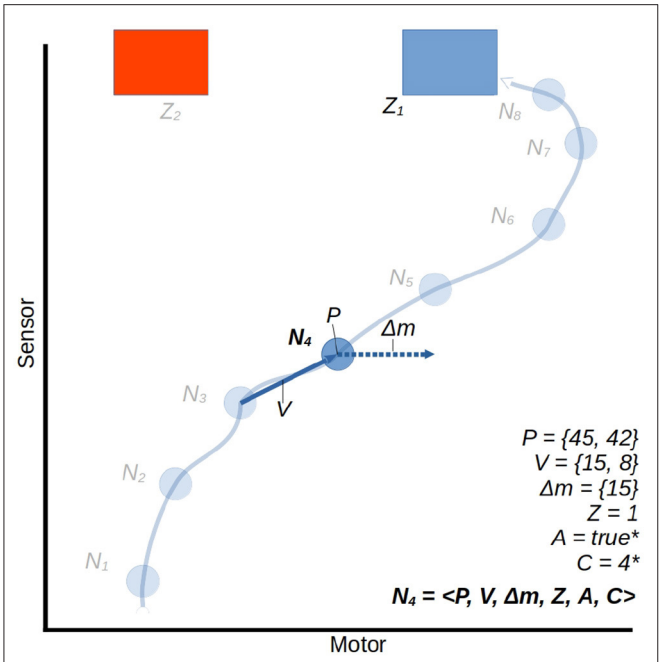



**FIGURE 6 |** A trajectory through a 2D sensorimotor space, represented in an ASM-unit with eight nodes. We use the node $N_4$ as an arbitrary example, all others share the same properties. For illustrative purposes we treat the motor and sensor dimensions abstractly as continuous ranges between 0 and 100. Note that the component $A = 1$ is determined in the context of subsequent activations of other ASM-units. This is explained in Section 2.4.

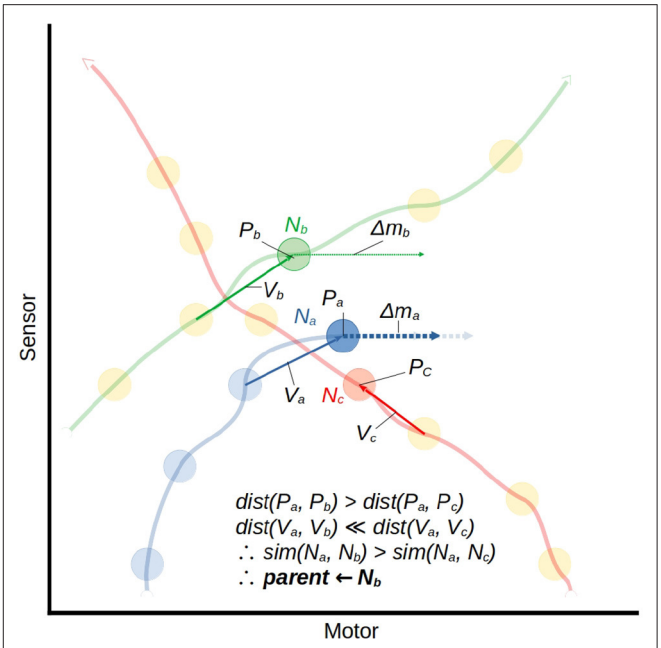



**FIGURE 7 |** A visualization of the mapping function using the similarity metric described in Equation (5). This illustrates the moment in which node $N_a$ is being created. The position in sm-space $P_a$ and displacement from the previous node $V_a$ will be compared to those of every nearby node. We isolate two of the three nearest nodes, $N_b$ and $N_c$, to compare. $N_c$ is closest in space to $N_a$, but the velocity of the trajectory associated with $N_c$ is very different from that of the current trajectory. $N_b$ is slightly further away, but the velocity of its associated trajectory is much more similar, so $N_b$ is selected as the parent node of $N_a$. $\Delta m_a$ is taken as the average between $\Delta m_b$ and the hypothetical vector which would put $N_a$'s successor at the same motor state as $N_b$'s successor. This is indicated by the pale arrows behind the $\Delta m_a$ arrow.

6. A boolean $A$ indicating whether the behavior associated with the node is reinforced ($A = 1$) or is inhibited ($A = 0$). (We will discuss this aspect of the model, which concerns its adaptive properties, in Section 2.4.).

A completed node may thus be defined as the tuple:

$$N = \langle P, V, \Delta m, C, Z, A \rangle \tag{4}$$

The number of nodes in each ASM-unit begins at zero and grows to a maximum of $N_{max}$ in a developed robot. After this maximum is reached, old nodes are destroyed to make room for new nodes. All of these nodes' data are used in future activations of the ASM-unit to contribute to future output of the mapping function. However, for now we will ignore the adaptive mechanism of the model and disregard the influence of the $C$, $Z$, and $A$ components of the nodes, which are involved in that mechanism. We will return to this aspect of the model in Section 2.4.

### 2.2.2. Motor Command Generation

At the same time as the generation of a new node to store state data, the model also generates a change-in-motor-state vector $\Delta$m which influences the current motor activity of the robot and is associated with the new node. This is done by finding a *parent node*, which is the historical node which represents a state most relevant to the current sensorimotor state of the robot. The parent node is found by a similarity metric which is applied to all historical nodes within a fixed distance of the current sensorimotor state in sensorimotor space, and the node which yields the greatest similarity value is classed as the current parent node. The behavior associated with the new node will be similar to the behavior associated with the parent node, and to reflect this the two nodes are regarded as having the same *class*. This is represented in-model with the new node's $C$ component set to the same value as the parent's.

The similarity metric which finds the parent node is illustrated by example in **Figure 7**. Let us consider a node which has just been created $N_a = \langle P_a, V_a, \Delta m_a, Z_a, A_a \rangle$ and an arbitrary historical node $N_b = \langle P_b, V_b, \Delta m_b, Z_b, A_b \rangle$ We measure the similarity of the historical node to the current node as the weighted product of the Manhattan distance between their positions in sensorimotor space and the distance between their incoming vectors.

$$sim(N_a, N_b) = -1 \left( \sum_{i=1}^{n} |P_{a_i} - P_{b_i}| \left( \sum_{i=1}^{n} |V_{a_i} - V_{b_i}| \right)^{\omega} \right) \tag{5}$$

Where $\omega$ is a fixed parametric weight which scales the relative importance of $V$ compared to $P$.

Once a parent node has been identified, it is used to determine a change in motor state for the robot. The method for this

is illustrated in **Figure 7**. The $\Delta m$ value for the new node is generated by taking a modulated form of the parent's $\Delta m$ value:

$$\Delta m_a = \frac{motor(P_a) - motor(P_b)}{2} + \Delta m_b \quad (6)$$

Where $motor(P_a)$ refers to taking only the motor components of the sensorimotor position. In effect this produces an interpolation of $\Delta m_b$ and the hypothetical vector which would yield $N_a$'s successor having the same motor state as $N_b$'s successor. If a parent node cannot be found, either because this is the first activation of this ASM-unit, or because no historical states were sufficiently close to the current state, then $\Delta m$ is generated randomly, with each component of the vector selected from a normal distribution ($\mu = 0, \sigma = 0.03$).

Once it has been generated, $\Delta m$ is used to determine a rate of change in motor state for robot over the next interval of $\tau$:

$$\dot{m} = \frac{\Delta m}{\tau} \quad (7)$$

In other words the robot's motor state changes linearly from time $t$ to time $t + \tau$ so that the motor state shifts from $m_t$ to $m_t + \Delta m$ over that interval.

## 2.3. Network-Level Model Architecture

We have discussed the design of an ASM-unit in isolation, and will now move on to how multiple ASM-units are linked together as a network as illustrated in **Figure 4**. As already mentioned, each ASM-unit spends only a limited period of time in a state of activation, and this state of activation regularly transitions from unit to unit. Transitions occur when the state of the system meets particular conditions, which depend upon either the robot's sensorimotor state or the duration of an activation. With a finite number of ASM-units in the network, walks through the network ultimately become cyclical, and this leads to the repeated activation of individual ASM-units which enable the history-based mapping functions to develop as they are applied.

The network, taken as a whole, defines the robot's behavior at a higher order than the immediate motor activations generated by an individual ASM-unit's mapping function: The complete activation of a specific ASM-unit reflects a directed transformation from one sensorimotor state to another (i.e., from one transition condition to another) over a discrete period of time, abstracted from the sensorimotor dynamics involved in producing that transition. In other words we may think of a complete activation of an ASM-unit as reflecting a performance of a discrete act (i.e., pushing a ball downwards), whereas the internal processes of each ASM-unit are reflective of the continuous sensorimotor dynamics that constitute that act (i.e., applying a certain amount of tension into the muscles as the surface of the skin feels a certain amount of pressure). Thus, similarly to the way that a set of sensorimotor trajectories captured in an individual ASM-unit reflect a set of regularities in a particular context of the agent-environment coupling, a repeated walk through the ASM-network reflects another set of historically-established regularities at a more coarse time scale. Having a multitude of ASM-units in the ASM-network produces a level of context-dependant and time-extended variability to the model's behavior: For any given sensorimotor state, one ASM-unit's mapping will likely give an output unique from any other ASM-unit. This means that the structure of the ASM-network, and the sequential order in which ASM-units are activated, is as fundamental to the behavior of the model as each independent mapping function.

In the general case, the ASM-network topology is dynamic and may generate new ASM-units over time and establish new links between ASM-units, however the details of this are not relevant to the investigation presented in Section 3 which uses a static network, and thus we will save that description for future work. Here we will focus on how groups of ASM-units are linked as a network and how transitions occur from one unit to another. **Figure 8** illustrates the different transition processes in a network. Each ASM-unit has a set of transition conditions $Z$, with

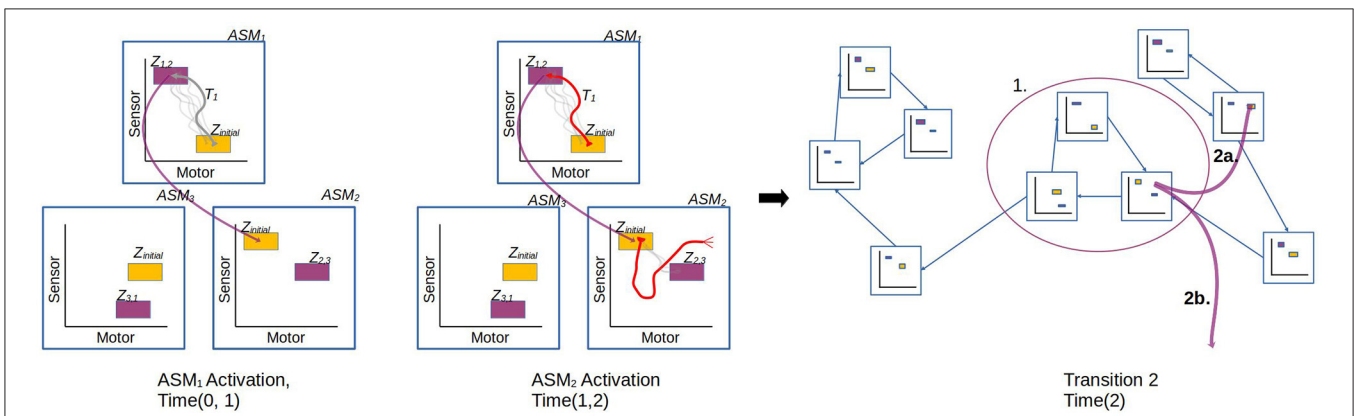



**FIGURE 8 |** Illustration of the transition process involved when an active ASM-unit succeeds ($ASM_1$) or fails ($ASM_2$) to establish the necessary sensorimotor conditions to make a transition that is associated with stable behavior. After $ASM_1$'s activation, the network transitions from $ASM_1$ to $ASM_2$, which have an historically established link. After $ASM_2$'s activation, however, the conditions to transition from $ASM_2$ to $ASM_3$ are not present. The controller explores SM-space until a stable behavior is re-established. It could do this by either by transitioning to an ASM-unit which was not previously linked (2a.), or through random motor activity (2b.).

each condition being associated with exactly one other ASM-unit. Each transition condition is defined as a hyperrectangular region in sensorimotor space with fixed upper and lower bounds along each dimension. Any time that the robot's sensorimotor state is inside one of these regions of the active ASM-unit, the transition condition is considered to be satisfied. When the condition is satisfied, activation of the current ASM-unit ceases and the ASM-unit associated with the transition condition becomes active. All ASM-units also define a set of initial conditions, which is simply the union of all of the transition conditions associated with that ASM-unit in *other* ASM-units in the network.

Transitions may also occur if the activation of a single ASM-unit has lasted for an over-extended period of time. The motivation for this is clarified in the next section based on the *principle of regularity* underlying the model's design. The limited time window is handled through a pair of parameters $t_g$, and $t_h$. The first defines a grace period, $t_g = 8$, in which only the active ASM-unit's transition conditions are checked. The second parameter defines a hard limit for the activation window, $t_h = 16$. In the times between $t_g$ and $t_h$ of an ASM-unit's activation, all of the network's initial conditions are checked as though they were transition conditions for the current ASM-unit. Finally, at time $t_h$, activation of the ASM-unit is terminated immediately and the controller generates random motor activity until any ASM-unit's initial condition is satisfied.

When a transition occurs, the formerly active ASM-unit's recently created nodes are updated with information about the transition condition, as explained in the earlier description of nodes. This aspect of the model is motivated by the need to adapt to irregularities in the agent-environment coupling, and other principles of structural self-individuation. We will now discuss the former concern in detail, but hold back discussion on the latter for a future work.

## 2.4. Adaptive Mechanisms of the Model

At last we turn to the adaptive mechanism of the ASM-network. This mechanism produces a simple intrinsic goal for every activation of an ASM-unit, toward which it is biased to develop: To establish both sensorimotor and environmental conditions that are sufficient to allow the next ASM-unit to do the same for its own successor, thereby maintaining the established structure of the ASM-network as a whole. To explain this, we begin by temporarily stepping back from the technical description to discuss how behavior can be understood as adaptive and maladaptive in the context of the model.

Recall **Figure 1**, which presented an illustration of sensorimotor scheme associated with bouncing a basketball. In that scheme, there is an established structure of regularities in the agent's movements and perceptions, and in the way that the ball responds to and enables them. However, if a disruption is introduced to the scheme, say the ball is the wrong shape to bounce in the same way as a basketball, performance of the scheme will quickly go away. **Figure 9A** illustrates such a disruption to the environmental response structure. In that example, the coordination in which the agent prepares to receive the returning ball is disrupted when the ball bounces away in a way that a basketball would not have. The previously established regularities in the relationship between motor action and sensory stimulation do not hold. The same actions associated with receiving the ball are met with irregular sensations, perhaps an

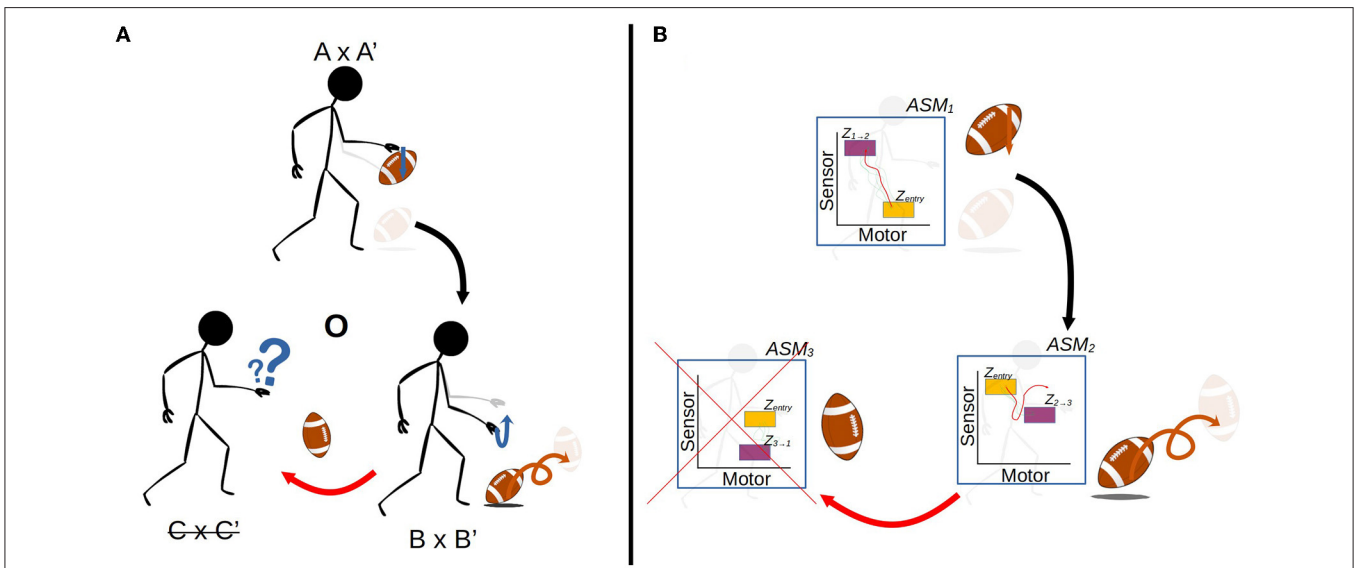



**FIGURE 9 | (A)** A visualization of a disruption to the basketball-bouncing scheme. The agent-side dynamics remain the same, but the environmental support for the scheme is insufficient, specifically in terms of the shape and other physical properties of the ball which cause it to bounce differently off the ground. This disruption prevents the transition from B × B' to C × C'. **(B)** An idealized illustration of an ASM-network controlling a robot encountering that disruption. Although the sensorimotor regularities involved in pushing the ball downwards are compatible which the established scheme, as the ball bounces wildly the sensorimotor relationship becomes irregular, perhaps due to the variation in the robot's visual sensors. This prevents $ASM_3$ from being activated as the robot is not in a suitable sensorimotor state.

emptiness of the hand and a sight of the ball moving away. The enabling conditions for the next coordination are not met, and the agent is at a loss. Successfully adapting to this disruption would entail altering the dynamics of the interaction with the ball such that the various normative conditions that motivate the agent to bounce the ball remain satisfied. At the sensorimotor level, this would mean enabling the continued performance of subsequent and concurrent sensorimotor coordinations as they have been established through experience.

In order to adapt to a disruption such as the misshapen ball, the agent could either adjust its movements so that it can bounce the different ball in such a way that it returns to hand, or it could do something other than bounce the ball if it is misshapen—perhaps kick it instead. In the former case, adaptation occurs within the context of a sensorimotor coordination—the dynamics involved in transforming the state of the coupling from one set of enabling conditions to another may alter while the same organization of coordinations is retained. In other words the agent may attempt to reconcile the disruption with the pursuit of the same goal. In the latter case, adaption occurs at the schematic level, through new coordination structures providing compensatory progressions through the same scheme, or with the emergence of a diverging sensorimotor scheme with a different normative orientation. These processes could also occur in tandem to a greater or lesser degree.

**Figure 9B** illustrates an idealization of the scenario of a disruption playing out in the case of a robot controlled by an ASM-network, comparable to **Figure 5** which illustrated the robot enacting the scheme without disruption. Just as the agent-environment coupling is not in a suitable state for enacting the next coordination, so is the ASM-network not meeting the conditions to allow the next ASM-unit to become active. Processes which compensate for such a disruption in ASM-network model could occur at both the unit and network level: Adaptation at the level of the coordination structure can be influenced through the reinforcement and inhibition of particular historical trajectories depending on how they relate to the resulting progression through the network. This alters the mapping functions and therefore the low-order dynamics of the coupling, while retaining the same higher-order sensorimotor transformations across sequences of ASM-unit activations. At the schematic level, adaptive processes can be influenced through the creation of new ASM-units in the network and new links between existing ASM-units. This allows new mappings to be generated and new transitions to occur, to accommodate new modes of agent-environment engagement. In this article, we focus purely on how the model's dynamics at the unit-level can adapt to maintain a pre-existing structure. The latter part of the adaptive process—how the structure of the network can generate dynamically—is equally important. However, we save that description for a future work as it is not a part of the investigation presented in Section 3.

The ASM-unit's adaptive mechanism is based on a *principle of regularity*. In a correctly functioning ASM-unit with an established set of historical trajectories, if the environmental state is sufficiently similar to its state during previous activations, then the sensorimotor trajectory produced by the ASM-unit's operation should also be similar to the trajectories produced by previous activations. By "similar trajectories" in this context we specifically mean two trajectories which begin within the same enabling conditions and reach the same set of transition conditions within a limited time window. This principle follows from the idea that because the controller is by design attracted toward repeating historical motor activity, the source of major deviations in a sensorimotor trajectory must be irregularities in the environmental response structure. Following from this principle, for a sequence of coordinations to be actively maintained over time, the stability of the environmental support structure must also be maintained. This provides a condition by which a sensorimotor trajectory may be evaluated in the context of enacting sensorimotor coordinations: Not only must there be regularity in the relationship between action and perception within a coordination, but that regularity must correlate with the stability of the environmental support for the next coordination. The model reinforces or inhibits trajectories based on whether that correlation appears to hold, based on the principle of regularity.

**Figure 10A** illustrates the process of reinforcing an instance of a behavior. The model always reinforces any behavior which does not lead to a failure to produce a regular trajectory in the next ASM-unit. In other words if the sensorimotor trajectory over the course of an ASM-unit's activation is similar to historical trajectories, then we assume that the environmental state delivered by the *preceding* ASM-unit activation provided suitable support for the sensorimotor coordination. It follows that the activation of the preceding ASM-unit did not establish any instability in the environmental support structure, and therefore that trajectory should be reinforced to have an attractive influence on future behavior.

In a contrasting example, **Figure 10B** illustrates an instance of nodes in a trajectory being inhibited because they lead to a breakdown of the established sensorimotor regularities. It follows from a corollary of the principle of regularity that if the current sensorimotor trajectory is different from historical trajectories produced by the same ASM-unit, then sufficient environmental support was not established by the *preceding* ASM-unit (e.g., $ASM_1$ in the figure) despite it achieving a suitable sensorimotor state. This means that the mapping associated with that ASM-unit, and the behavioral dynamics that it produces, are not sufficient to maintain the stability of the broader sensorimotor structure, because it produced regularities in the sensorimotor response that did not correlate with the stability of the environmental support structure. Therefore, the nodes associated with the dynamics of the preceding ASM-unit's last activation are inhibited, in order to alter the unstable dynamics. Additionally, the nodes in the current ASM-unit (e.g., $ASM_2$ in the figure) which fails to transition to the expected next ASM-unit are also inhibited, as they too reflect dynamics which did not produce a stable engagement.

Let us return to the details of the implementation. Trajectories are reinforced or inhibited by setting the value of $A$ component of every node involved in representing that trajectory. Recall that this component is simply a marker of this reinforcement property: if the trajectory is reinforced, then $A = 1$ for every node

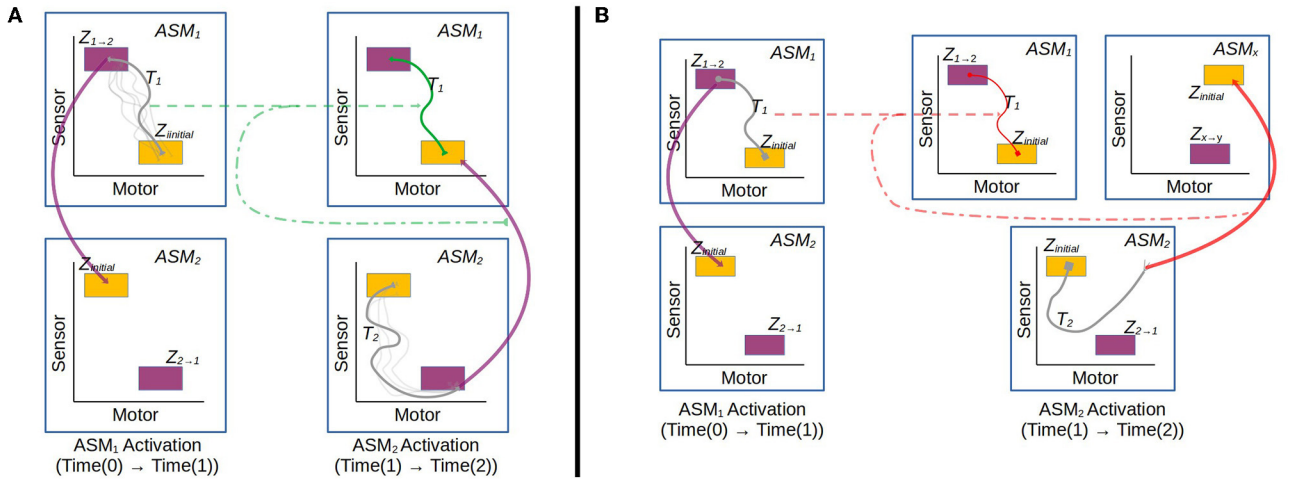



**FIGURE 10 | (A)** Illustrates an instance of reinforcement. At the end of Activation 2 there is a successful transition from $ASM_2$ to $ASM_1$, and this leads to a reinforcement of the trajectory from Activation 1 which established the conditions for that successful transition. **(B)** By contrast illustrates an instance of inhibition. At the end of Activation 2 the established conditions to transition to $ASM_1$ are not met, and activation passes to some other unit $ASM_x$. This causes the trajectory from Activation 1 to be inhibited because it produced conditions which led to an unsuccessful transition.

associated with that trajectory. If inhibited, then $A = 0$ for every node. Our explanation of the ASM-unit's mapping function in Section 2.2 assumed that all nodes were reinforced, but we now complete the explanation in the case where nodes may be either reinforced or inhibited. Recall that we previously stated that if $A = 1$ for all nodes, then the parent node is identified as the historical node which yields the highest similarity score in the metric given in Equation (5). When the adaptive component is included however, and $A = 0$ in some cases, the ASM-unit uses a filtering process to bias the system toward repeating behavior associated with the most relevant reinforced historical trajectory, even if there are several other more similar historical trajectories. The process may be best described algorithmically:

1. $sim(N_a, N_i)$ (Equation 5) is applied to all nodes to find the node which produces the highest similarity score, call it $N_b$.
2. If $N_b^A = 1$, then $N_b$ is regarded as $N_a$'s parent node and the algorithm terminates.
3. Otherwise, a set $\mathbf{C}$ of node class labels is created such that $\mathbf{C} = \{N_b^C\}$.
4. The node with the *next* highest similarity score is found, call it $N_c$.
5. If $N_c^A = 1$ and $N_c^C \notin \mathbf{C}$ and $N_c^Z \neq N_b^Z$, then $N_c$ is regarded as $N_a$'s parent node and the algorithm terminates.
6. Otherwise, set $\mathbf{C} = \mathbf{C} \cup \{N_c^C\}$ and return to step 3 until the algorithm terminates or there are no more valid historical nodes for comparison.

Once the algorithm terminates, the $\Delta$m change-in-motor-state is generated as described earlier. This process causes the behavior of the robot to be directed toward repeating dynamics which ultimately supported successful transitions through the network, as well as actively avoiding those which failed. Essentially the addition of new, reinforced nodes representing an instance of a behavior increase the likelihood of that behavior's future performance by (1) increasing the diversity of states which attract the repetition of that behavior, and (2) by lasting longer than the nodes which came before them, given the finite capacity for nodes in an ASM-unit. By contrast, the addition of inhibited nodes reduces the likelihood of the same behavior being repeated in the future by negating the influence of reinforced nodes *via* the effect of the $C$ and $Z$ components.

This completes our description of the model as used in this investigation. The reinforcement and inhibition mechanisms produce a simple intrinsic goal for every activation of an ASM-unit, toward which it is biased to develop: To establish both sensorimotor and environmental conditions that are sufficient to allow the next ASM-unit to do the same for its own successor. We now present results of an investigation which demonstrate how this intrinsic goal can in turn produce more second-order goal-directed behavior in the robot which the ASM-network controls.

## 3. INVESTIGATION

We now demonstrate how an ASM-network can be used to control a robot which successfully learns to perform a task involving object discrimination. The parameters of the robot and environment are essentially equivalent to an experiment first presented by Beer (1996), in which agents were evolved to distinguish between circles and diamonds using the standard evolutionary robotics technique of evolving a continuous-time recurrent neural network controller using a genetic algorithm. The agents demonstrated their ability to distinguish between the shapes by colliding with circles while avoiding the diamonds. The task captures a fundamental capacity of any acting agent—in order to selectively interact with its environment, an agent must be capable of discriminating between different environmental

features. We use this task as a first demonstration of the ASM-network's value in investigating goal-oriented adaptive behavior guided by the relationship between environmental and internal mechanisms, as opposed to extrinsic fitness functions.

In our version of the experiment no genetic algorithm or any other external optimization process is required. The properties of the environment are specifically arranged, and the ASM-network model is partially constrained, so that the stability conditions for the ASM-network's dynamics concurrently produce behavior which aligns with the ascribed norms of the task. Initially the robot's behavior is entirely random, but over time the robot develops an ability to scan the shape, identify the difference between circles and diamonds, and responds appropriately to the different shapes. Our results illustrate how our robots solve the task.

## 3.1. Experimental Setup

**Figure 11** illustrates the experimental setup. A robot with seven ray sensors and one bi-directional motor is situated in a 2 dimensional arena. The rays are spread evenly with an angle of $\frac{\pi}{9}$ radians between each, with three on either side of a central ray pointing directly upwards. The motor allows the robot to move horizontally with a velocity ranging between −30 and 30 units per second. The arena has a width of 300 units and a height of 300 units, and periodic boundaries. At the start of a simulation the robot begins at position (150, 0) in this arena. An object, which may initially be either a circle or a diamond shape is positioned in the arena. The object enters the arena at 100 units above the robot vertically and offset between −50 and 50 units horizontally from the robot. Circle objects have a diameter of 36 units, and diamonds have a side length of 36. The objects falls at a rate randomly selected between 12 and 16 units per second. The robot's sensors are stimulated whenever the ray intersects with the falling object, with the sensor activation modeled as a continuous scalar which linearly increases from 0 if the intersection point is at the tip of the ray, up to 1 if the intersection is at the position of the robot.

A single run of the experiment continues until there have been at least 2,000 descents of both circles and diamond objects. The results here are based on 64 runs. During a run, the object falls directly downwards, while the robot moves around the

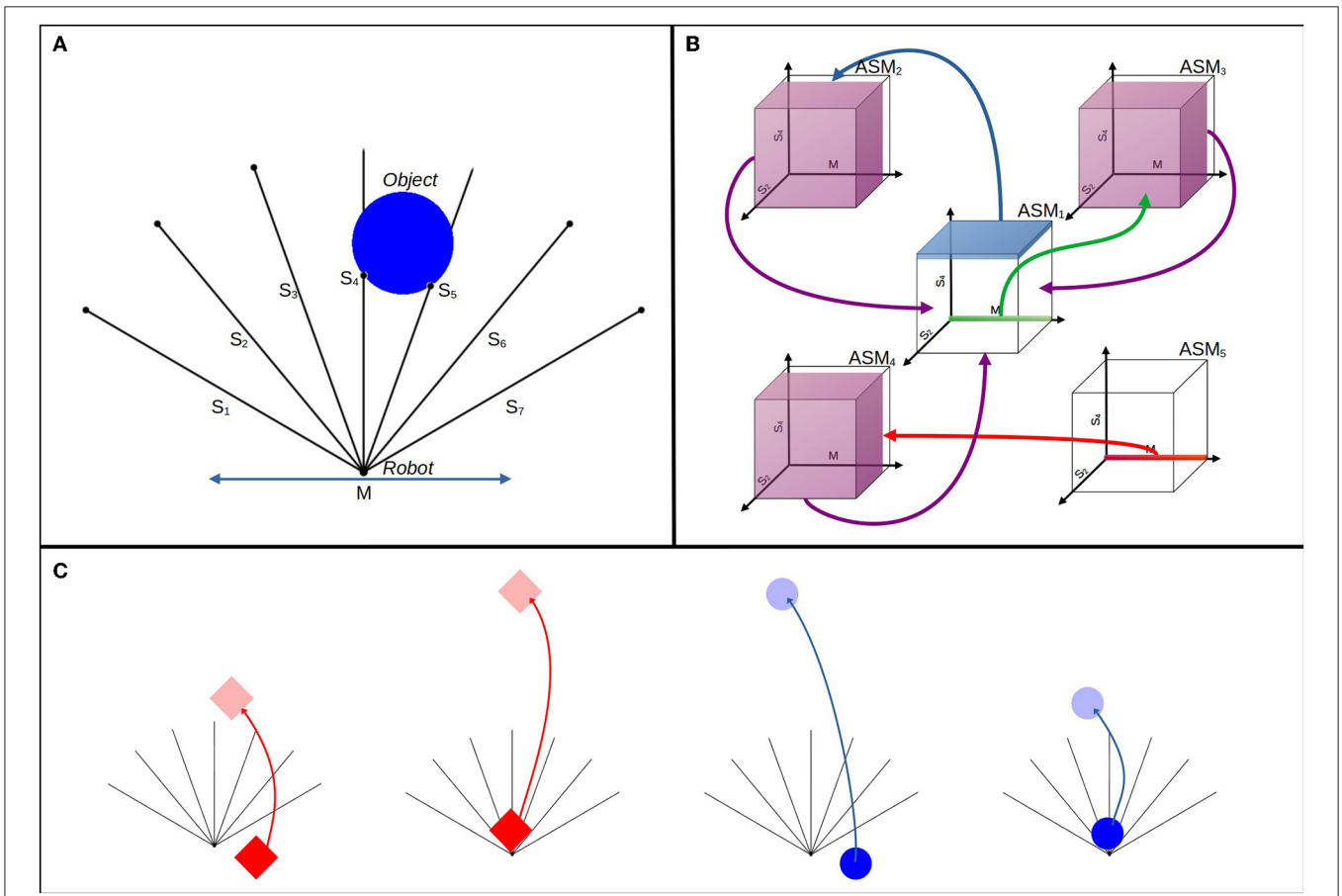



**FIGURE 11 |** Experimental setup in three parts. **(A)** Illustrates the arrangement of the robot, its sensors, and an example of the circle descending. **(B)** Illustrates the topology of the network that controls the robot. Transition conditions are suggested in three dimensions, with arrows and color-coding indicating which ASM-unit is enabled by each transition condition. These conditions are discussed more precisely in the body text. **(C)** Illustrates the different ways that the object types react to hitting the robot or the arena floor.

arena freely. When the object collides with the robot or the bottom of the arena, it returns back toward the top of the arena immediately, responding differently if it is a circle or a diamond. This is illustrated in **Figure 11C**. If a circle collides with the robot, it returns 100 units vertically, whereas if it collides with the bottom of the arena it returns 300 units. The inverse is true for the diamond, it returns 100 units when it hits the bottom of the arena and 300 units when it collides with the robot. When the object returns, its downward velocity randomly resets to a new value from the same possible range, the object's horizontal offset is randomly reset, and the shape has a chance of swapping to the other type with a probability of $p = 0.5$. This closely resembles a succession of resetting trials in an evolutionary robotics framework, but we emphasize that a single robot remains active over the course of the entire run, and the single ASM-network develops its history over the run. The continuity of a run is critical for two reasons: (1) The ASM-network needs to build a history of behavior over time for it to develop toward solving the task; (2) There is a consequential difference for the robot between colliding with an object and missing an object, in that the object returns to its sight more or less quickly depending on its shape and whether it collides or not (i.e., diamonds returning either 300 or 100 units up and the opposite property for the circles). The significance of these differences is discussed in detail in Section 4.

Each ASM-unit in the network has a sensorimotor space with eight sensorimotor dimensions ($S_{1..7} \in [0,1]$, $M \in [0,1]$), coinciding with the seven sensors and one motor ($R^{S_{1..7}} \in [0,1]$, $R^M \in [-30,30]$) of the robot, such that:

$$M = \frac{R^M + 30}{60} \tag{8}$$

$$S_x = R^{S_x} \tag{9}$$

The network is constrained to support the fulfillment of the task. The network has a fixed arrangement of five units, with pre-given transition conditions and links between each. These transition conditions are associated with the potential sensorimotor states of particular stages of the desired functional behavior, e.g., when the robot sees any object, when the robot collides with any object. This scaffolds the development of functional behavior and constrains which habits are potentially viable, but it does not define the behavior of the robot, as all of the motor dynamics are produced by the ASM-unit mapping functions, which begin undefined as there is no history for them to respond to. The way in which these constraints scaffold specific functional behavior is explained in Section 4.

The topological arrangement of the network is illustrated in **Figure 11B**, but due to the dimensionality of the sensorimotor space the transition conditions are only able to be suggested in an image. We define them precisely here. $ASM_1$ has 2 separate transition conditions linked to $ASM_2$ and $ASM_3$, respectively. The transition condition $Z_{1,2}$ (i.e., condition for the transition from $ASM_1$ to $ASM_2$) is defined as follows:

$$Z_{1,2} : \left\{M \in [0,1], S_{1,2,3,5,6,7} \in [0,1], S_4 \in [0.98,1]\right\} \tag{10}$$

Where $M \in [0,1]$ means that the motor state as represented in the ASM-unit may be anywhere between 0 and 1 to satisfy the condition. $S_x$ refers to the same for each sensor. Note in particular that $S_4$ is different from the others. Practically, this means that the transition occurs whenever the robot's central sensor is very highly stimulated, and all other sensors and the motors may be in any state. This condition would occur whenever the object collides into the front of the robot. The other transition condition is:

$$Z_{1,3} : \left\{M \in [0,1], S_{1,2,3,4,5,6,7} = 0\right\} \tag{11}$$

Which means that this condition is satisfied if and only if every sensor is at 0, i.e., the robot cannot detect the object.

$ASM_2$ has two transition conditions which are both linked back to $ASM_1$:

$$\begin{aligned} Z^a_{2,1} &: \left\{M \in [0,1], S_{1,3,4,5,6,7} \in [0,1], S_2 \in [0.01,1]\right\} \\ Z^b_{2,1} &: \left\{M \in [0,1], S_{1,2,3,4,5,7} \in [0,1], S_6 \in [0.01,1]\right\} \end{aligned} \tag{12}$$

Which means that the conditions are satisfied if either the $S_2$ or $S_6$ sensors are at least slightly activated. In practice, at least one of these conditions is satisfied if the object is anywhere in the majority of the coverage of robot's sensory field, although not if for instance the object is moderately far to the left or right, or immediately in front of the robot at a long distance.

$ASM_3$ has two transition conditions which are the same as those in $ASM_2$, such that $Z^a_{3,1} \approx Z^a_{2,1}$ and $Z^b_{3,1} \approx Z^b_{2,1}$. We use the approximation to reflect that although the ranges are the same, the transition conditions are not identical because the sets require different ASM units to be active. $ASM_4$'s transition conditions are also defined similarly to other units, such that $Z^a_{4,1} \approx Z^a_{2,1}$ and $Z^b_{4,1} \approx Z^b_{2,1}$, and finally $ASM_5$'s transition condition is $Z_{5,4} \approx Z_{1,3}$. In Section 4, we discuss how this arrangement relates to the behavior that the controller produces in more detail.

## 3.2. Results

We measure the performance of a robot in the task by looking at how many times the robot responded "correctly" in a window of the most recent descents of each shape. **Figure 12** illustrates the average performance over time over 64 runs. The plot samples every 20th descent of either circles or diamonds, with each point giving $\sum_i \frac{c_i}{20\times 64}$, where $c$ is the number of correct responses in the previous 20 descents of a shape for the $i$th robot. Across 64 runs, the average performance of the robots for the first 20 descents for circles is 0.37 (i.e., they catches circles 37% of the time) and for the first 20 descents of the diamond is 0.76 (i.e., they avoid diamonds 76% of the time). The average performance in the last 20 descents for circles is 0.97, and for the last 20 descents of diamonds the performance is 0.99. Performance improvement is rapid, reaching an average of over 0.9 for both shapes within 200 descents, and reaching peak performance after 1,000 descents. Performance for diamonds is higher, especially at the start, because of the greater likelihood of missing an object by chance compared to colliding with that object. Catching involves precise positioning, whereas avoiding can be accomplished in many equally good ways. These results illustrate that the robot

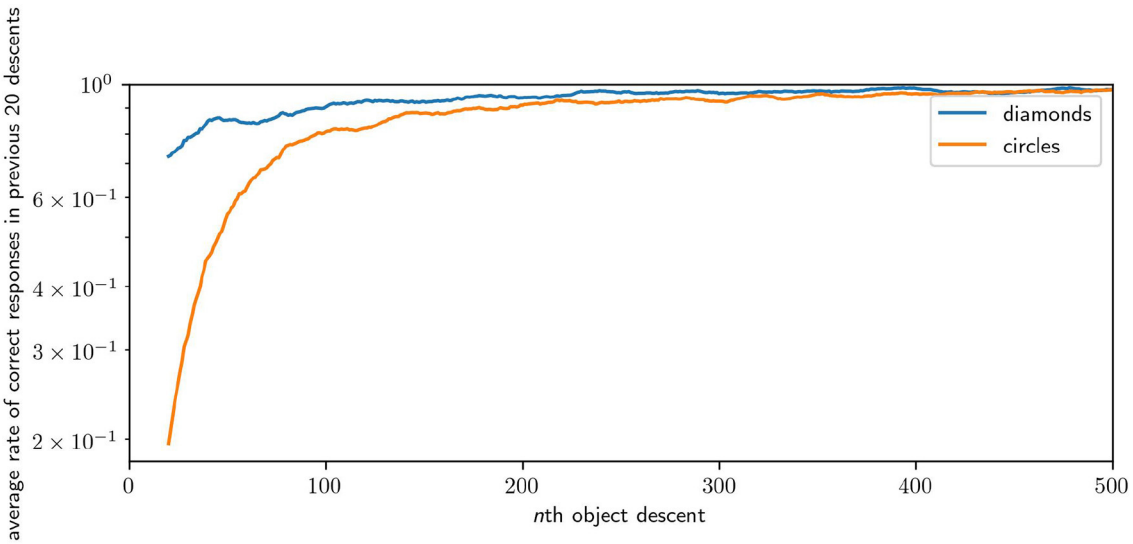



**FIGURE 12 |** Plot of the improving performance of robots, averaged across 64 runs. Although this figure only shows the first 500 interactions with each shape because most of the development occurs early, performance continued to improve incrementally for the remaining 1,500 interactions.

is capable of learning how to effectively discriminate between diamonds and circles.

The robot is capable of reliably solving this specific task because the topology of the network is arranged such that the behavior involved in maintaining the established links in the network is necessarily also behavior which solves the task. In all robots we observed a direct correlation between the rate of successful transitions between ASM-units and the rate of the robot's correct responses to the object shapes. Crucially though, the pre-given topology and environmental conditions are insufficient to define for the robot the actual sensorimotor dynamics involved in solving the task, i.e., how to move around the environment in such a way that it can identify the different shapes and collide or avoid as appropriate. To learn these dynamics, the robot must engage with the environment over time, and over the course of this engagement the model's adaptive mechanisms reinforce those dynamics that support the transition conditions within the network and inhibit those that lead to violations of those conditions. The maps of each ASM-unit, most critically $ASM_1$, develop in such a way that the robot's behavior consistently establishes the enabling conditions of each ASM-unit in a suitable sequential order. In $ASM_1$ this means that the mapping must produce dynamics which differ when encountering differently shaped objects, as the environmental conditions have been set such that disruptions will occur elsewhere in the network if the robot interacts with the objects incorrectly. This means that the bulk of the learning process that occurs over the course of an experimental run is in the development of $ASM_1$'s map.

Meaningfully visualizing the maps themselves, and how they change over time, is challenging due to their dimensionality. **Figure 13** illustrates a projection of the states of $ASM_1$ for one robot using principle component analysis. Note that Component 1 $= [0, -0.03, -0.03, 0.04, 0.18, 0.44, 0.64, 0.59]$ and Component 2 $= [0, -0.01, -0.01, 0.20, 0.66, 0.56, -0.32, -0.29]$. The plot compares node positions in the early and late stages of the robot's development and highlights important differences. Firstly, almost all nodes are reinforced by the end of the robot's development. Secondly, there are subtle change in the distribution of the nodes over time, and by extension the mapping: One clear example is that inhibited nodes (purple) tend to be clumped together around $(0.73, 0.9)$, in the early stages of development, and by the later stage of development most of the nodes in that area have disappeared. This suggests that the inhibited nodes successfully dissuade the continuation of trajectories which approach that region of sensorimotor space. **Figure 13C** illustrates the projected sensorimotor trajectories of the last 10 interactions with each shape for the same robot, indicating the way that the trajectories for different shapes diverge around $(0, -0.25)$, and end in different regions of sensorimotor space.

The relative positions of the robot and object over time as the object descends through the robot's sensory field provide a better illustration of the robots' development. **Figures 14**, **15** illustrate examples of these trajectories. **Figure 14** demonstrates the development of a single robot, which we selected to exemplify the way in which the adaptive mechanisms of the model contribute to particular dynamics becoming more stable than others depending on how well they support the network-level organization. The plots represent the robot's horizontal position relative to the object, from the moment the object is at a height of 100 units until the moment the object reaches height of 0. In each encounter the initial conditions of the object vary, in terms of speed and displacement, within the parameters already discussed. The top row illustrates the first 200 encounters with a circle, and then the last 50 encounters. The second row illustrates

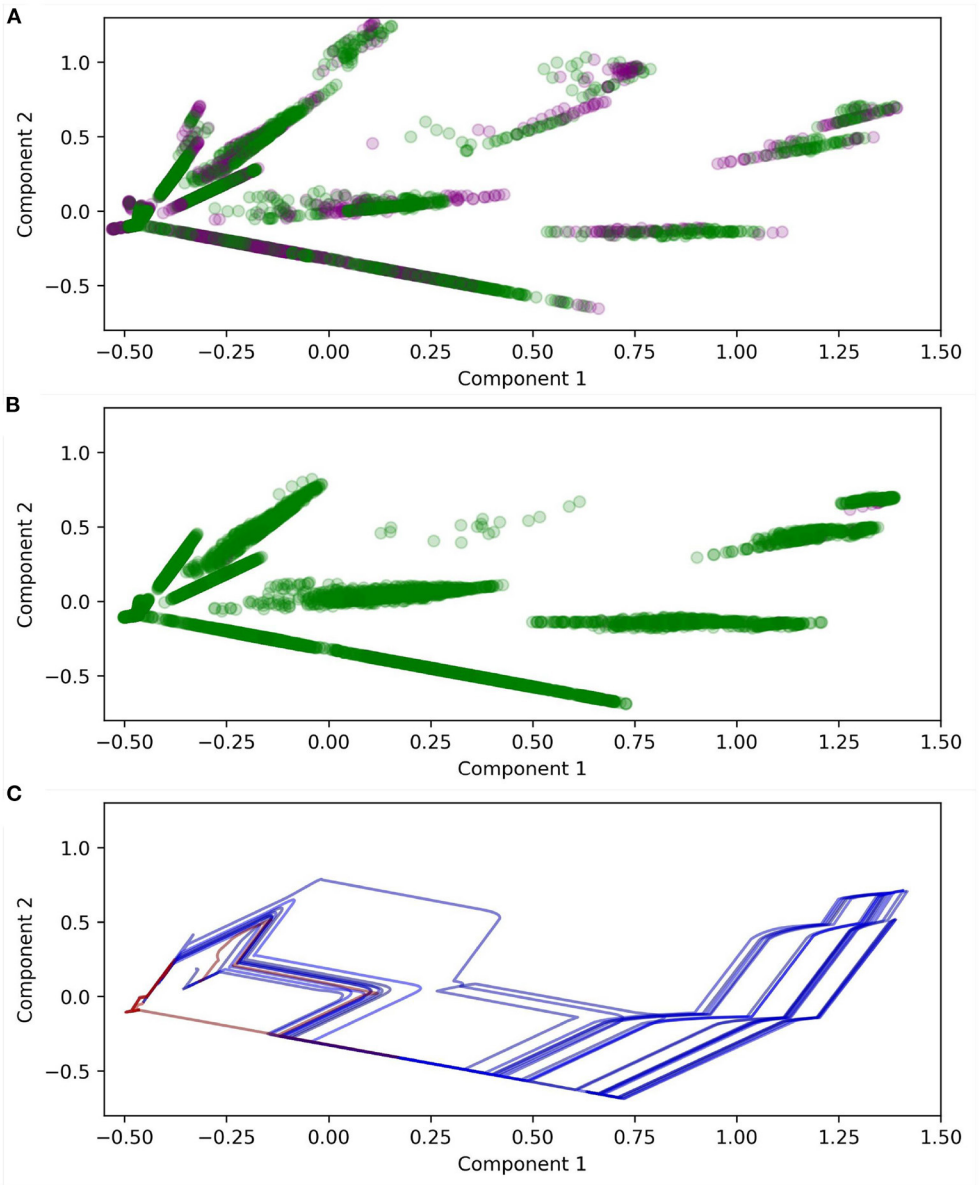



**FIGURE 13 |** Projections of $ASM_1$ derived using principle component analysis. **(A)** Illustrates the positions of the nodes after 100 descents of the object. Purple markers indicate inhibited nodes, green indicates reinforced nodes. **(B)** Illustrates the positions of nodes after 2,000 descents. **(C)** Illustrates the sensorimotor trajectories of the interactions with the last 10 descents of each shape. Blue indicates circles and red indicates diamonds. Note that the motor dimension is ignored because it confounds the interpretability of the plot. The rapid jumps in the plots are due to individual sensors suddenly becoming active or inactive as they intersect with the object.

the same for encounters with the diamond. The majority of the improvement occurs rapidly, in the first 50 encounters. We can also see that more subtle developments continue, most notably a kind of behavior which leads to occasional narrow misses of the circle becomes less frequent between the 50th to 200th encounter. The final trajectories shows how the developmental process has continued over the longer term to exaggerate the differences between responses to the different object shapes, and to increase the consistency of responses to a particular shape, especially in the case of diamond encounters.

**Figure 15** demonstrates the performance of eight randomly selected robots, contrasting the early stages of their interactions with the objects against the late stages. Every robot displays a tendency to transition from an initially sprawling set of different trajectories to a significantly more concentrated set of trajectories in the later stages. Given that the object appears at a random position with respect to the robot, a typical strategy emerges which involves the robot moving to approximately the same position relative to the object in each encounter, before the responses to the different shapes diverge. All of the robots display

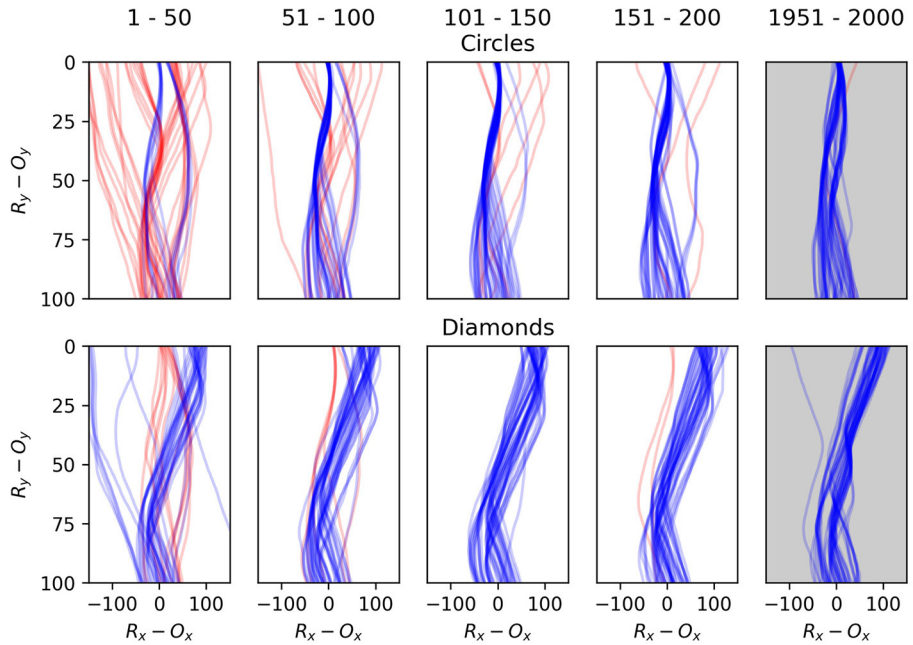



**FIGURE 14 |** Plots illustrating the behavior one robot as it interacts with the falling objects. The first column shows the robot's first 50 interactions with circles in the top row and diamonds in the second row. The second column shows the second 50 interactions and so on. Each line illustrates a single interaction with an object from the moment that it is 100 units above the robot until the moment the object either collides with the robot or the bottom of the environment. Blue lines indicate that the object ultimately did the correct thing as per the task description, i.e., it collides with the circle or avoids the diamond. Following each plot from bottom to top shows the relative horizontal positions of the robot and object as the object descends toward the robot. The object's position is always at 0 on the x-axis, so a point at x = 100 indicates that the robot is 100 units to the right of the object, and gives no indication of the robot's absolute position in the environment.

some variation of a behavior involving sweeping multiple sensors across the object in either direction before the responses diverge. Beer observed that similar foveate-scan-decide strategies were typical in his evolved CTRNN-controlled robots. This suggests that these kinds of responses are particularly attractive for this task even when the adaptive mechanisms producing such behavior are distinct.

## 3.3. Auxiliary Results

In an auxiliary experiment we, performed 16 runs in which robots were only ever exposed to diamond objects. There we observed that a dominant behavior is for the robot to immediately and continually moving at full speed in one direction or the other, thereby missing the diamond by some distance. This is a very simple behavior for the robot to discover by accident as it simply involves keeping its motor state around its maximum or minimum regardless of sensory state. This behavior also appears in the early stage of **Figures 15A,D–F**, but is lost by the later stages. This suggests that such a behavior is less stable when circles are also present, as it limits the robot's ability to identify the shape type and respond appropriately. This provides an example of the space of viable habits being constrained by the contrasting properties of the shapes. Finally, we performed 64 runs in which the responsive properties of diamonds and circles was inverted. The average successful performance rate over the course of all robots' development is presented in **Figure 16**.

# 4. DISCUSSION

## 4.1. How Can the Robots Perform the Task?

Our results present a model, based on enactive principles of sensorimotor contingency theory and habit, which allows a robot to learn to perform a specific cognitive task of object discrimination without a functionally-oriented reward mechanism. The model as presented in Section 2 is a generic medium which specifies a whole suite of dynamics with certain kinds of attractors. In the experiment presented in Section 3, we apply some constraints to that medium which establish the viability conditions of a particular sensorimotor organization, such that the internal norms of the system align with the ascribed norms of task, that is to avoid diamonds and collide with circles. This allows a generic adaptive mechanism, directed toward satisfying those internal norms, to also shape the behavior of the robot to satisfy the requirements of the task. In natural systems, a web of evolutionary and developmental processes all serve to shape the sensorimotor organization of an agent in a manner that produces an alignment such as this, while we have engineered the alignment with a specific set of constraints utilizing our knowledge of the task and system. How exactly does the process of maintaining this particular sensorimotor structure align with adaptively regulating behavior in terms of a functional task?

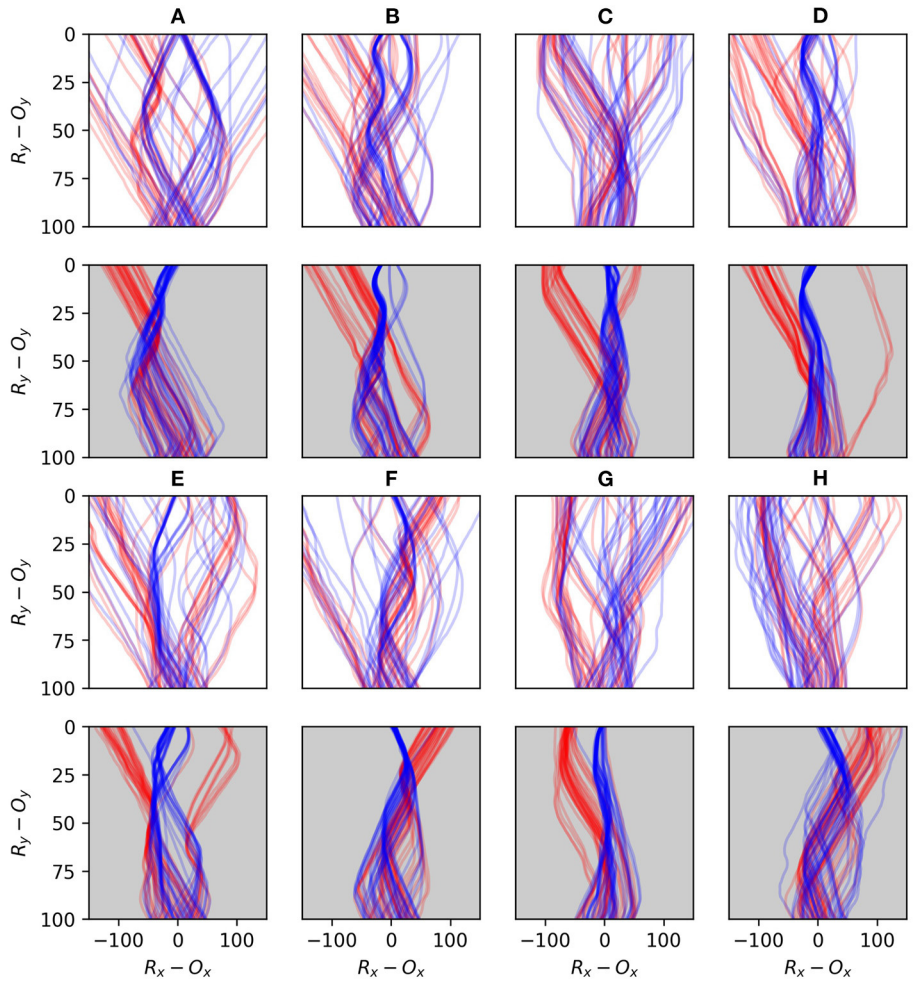



**FIGURE 15 |** Plots comparing the early and late stages of development for eight randomly selected runs. The first and third rows illustrate the first 40 interactions with each shape type for 8 different robots **(A–H)**, and the second and fourth rows show the last 40 interactions for the same robots. Blue lines indicate interactions with circles and red lines interactions with diamonds.

**Figure 17** illustrates the relationship between the enabling conditions of an organization of sensorimotor coordinations as they relate to the structure of the ASM-network used in the experiment. We conceptualize an agent performing this task as alternately enacting a pair of habitual behaviors, which, following (Egbert and Barandiaran, 2014), we understand as simple loops of sensorimotor coordinations:

1. In Loop 1 the interaction progresses from the robot detecting the object at long range until it collides with the object ($A_1 \times A_1$'), and then from there until the robot detects the object at long range once more (B × B').
2. In Loop 2 the interaction progresses from detecting the object at long range until the object leaves the robot's sensor range ($A_2 \times A_2$'), and then until the robot detects the object at long range once more (C × C')

We label $A_1$ and $A_2$ as such because they share an enabling condition but involve diverging sensorimotor trajectories to reach different transition conditions. Thus, we have two partially overlapping loops of sensorimotor coordinations $A_1 \times A_1$' → B × B' → $A_1 \times A_1$' and $A_2 \times A_2$' → C × C' → $A_2 \times A_2$. As we have discussed in Section 1, we may think of the continuing sequential satisfaction of the enabling conditions in these loops as the conditions of viability of a sensorimotor habit. By design, our ASM-network medium is oriented toward adapting behavior to maintain such viability conditions. Transition conditions between $ASM_1$, $ASM_2$, and $ASM_3$ are associated with sensorimotor states associated with detecting objects, colliding with objects, and losing detection of objects, while $ASM_4$ is associated with conditions that only occur when the expected progress through the loops is disrupted. Thus the pre-given structure of the ASM-network in **Figure 11** imposes this kind of arrangement of sensorimotor coordinations onto the robots, and as such the sequential fulfillment of the coordinations' enabling conditions as its task. However, those habitual viability conditions are completely

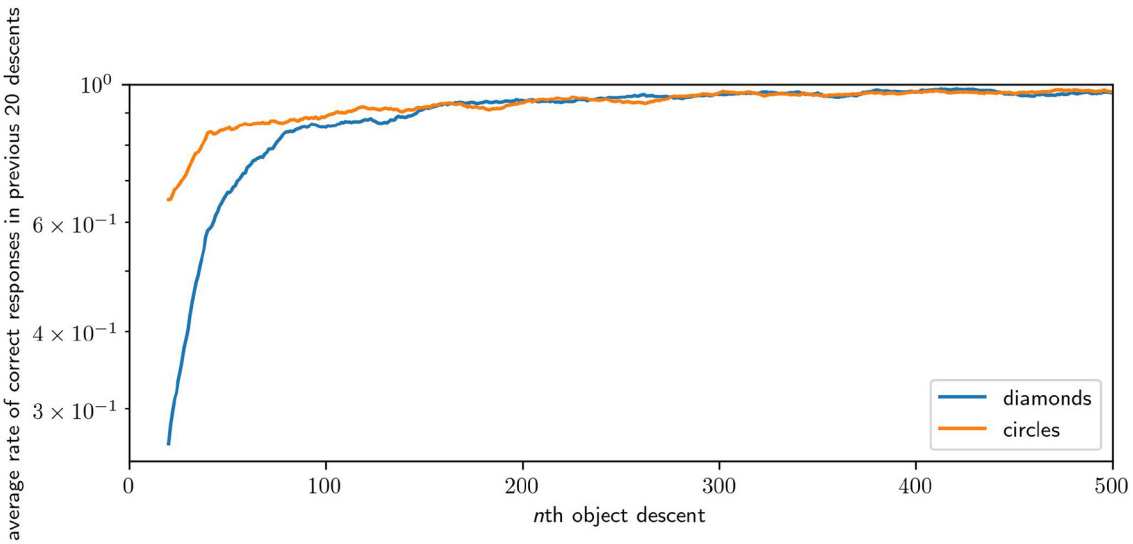



**FIGURE 16 |** Results of an inversion of the default task, in which the responses properties of diamonds and circles are swapped. This causes the robots to develop to avoid circles instead of diamonds and *vice versa*. Results are presented as per **Figure 12**, note that the learning rates are equivalent but the robots take longer to respond correctly to diamonds instead of circles in this case.

agnostic to the different sensorimotor properties of interacting with the shapes, and thus do not sufficiently explain the robots' functional fulfillment of the object discrimination task. It is the different properties of diamonds and circles, with respect to what happens when they collide with either the robot or the bottom of the arena, that imbues the shapes with intrinsic relevance with respect to these goals. Specifically, colliding with a diamond causes a delay in returning the object which disrupts the progress of Loop 1, and likewise missing a circle leads to a disruption Loop 2. Since the ASM-network's adaptive mechanism is geared toward avoiding behavior which produces disruptions, i.e., behaviors which are non-viable with respect to maintaining the arrangement of coordinations, the difference between the objects will drive the robot to respond differently to the two shapes. An interesting consequence of this is that if we invert the properties of circles and diamonds in terms of how they respond to collisions and misses, then robots with the same ASM-network parameters instead learn to seek diamonds and avoid circles. That the functional behavior produced is a equally a consequence of both the internal dynamics of the agent and dynamic properties of the environment highlight the value of this kind of experimental approach.

## 4.2. What Do the Robots Learn Autonomously?

We have discussed how the particular network used in the experiment produces an alignment between the internal mechanisms of our robot and the ascribed norms of the task. But since we achieve this alignment through directly engineering a set of constraints, what exactly is left to the robot to learn autonomously? In functional terms, the robot has only been given a structure of sensorimotor conditions that it needs to repeatedly satisfy in order to maintain stable behavior. It must learn that there is a difference in the way that the two shapes impact that stability, that the difference corresponds with particular perceptual characteristics of interacting with the shapes as they descend, and how to act in response to those different characteristics so that it avoids interactions which destabilize its sensorimotor structure. While the arrangement in **Figure 17** are a consequence of the network structure illustrated in **Figure 11B**, the development of suitable ASM-unit mappings which satisfy this arrangement is comparable to the optimization of weightings in terms of task fitness in an Evolutionary Robotics approach.

We can explore this further to clarify our model's relationship to the theoretical concepts of sensorimotor contingencies and habit mentioned in Section 1. Our constraints on the network establish the parameters of the relationships between a set of sensorimotor coordinations that are necessary for those coordinations to be stable, but it does not establish the actual sensorimotor dynamics that constitute those coordinations. While the general effect of action on perception is implicitly established in the characteristics of the robot and shapes, closing the causal loop to establish the effect of perception on action can only be established through interaction between robot and environment. The development of those agent-side dynamics will vary based on whatever specific environment-side dynamics it encounters. The environmental dynamics can vary in terms of four properties, all of which change the perceptual character of interacting with the objects from the agent's perspective. These properties are: (1) the different shapes of the objects; (2) The different mechanical properties of the objects (i.e., what happens upon collision); (3) The different speeds of the objects; (4) The different initial displacements of the objects. With our

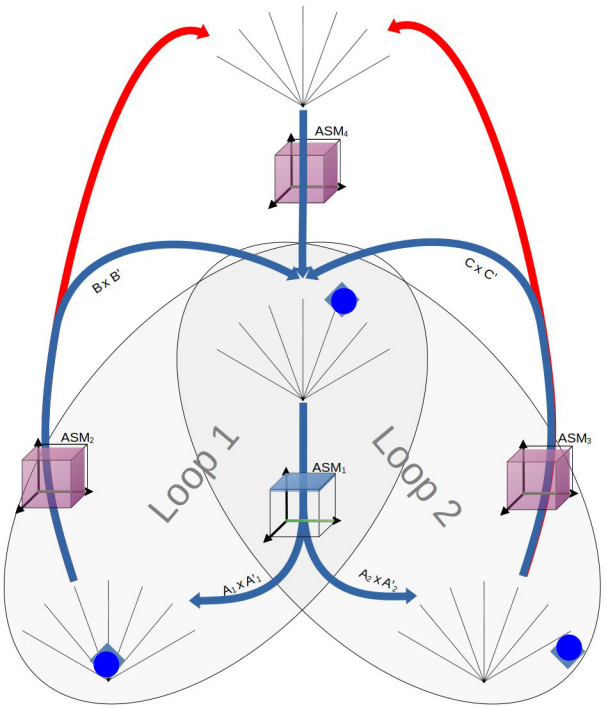



**FIGURE 17 |** An illustration of the robot's arrangement of sensorimotor coordinations as they relate to the ASM-network. Successful performance of the task aligns with the robot proceeding through performance of Loops 1 and 2 without disruption. Over time $ASM_1$ develops such that only $Transition_3$ occurs when diamonds are present and only $Transition_2$ occurs when circles are present. Some simple adaptive processes also need to take place in $ASM_2$ and $ASM_3$ such that they consistently lead to the robot re-discovering the object as it returns into view. Note that $ASM_5$ does not feature as its only purpose in the network is to provide suitable initial conditions for $ASM_4$.

privileged view of the system we know that the differences in properties 1 and 2 are meaningful with respect to the norms of the system, while 3 and 4 are not, but this is not made explicit in the experimental setup, i.e., the adaptive mechanism is not tuned to respond to those properties in the same way that a fitness function defines the relevant properties of the world. Through interacting with the environment over time, these relevancies nevertheless become expressed through the robot's behavior.

In developing stable habitual behavior, the robot effectively learns that it needs to respond to the different shapes differently and how to make those different responses. While it is not responding like this, the internal dynamics of the controller will be in flux since the agent-side dynamics of the coupling become altered when particular trajectories are inhibited. Stable habitual behavior entails the performance of a habit continually re-establishing the conditions of its own re-performance, but due to the instability of the controller the conditions for a particular way of performing a habit may be lost over time even if the same initial sensorimotor state is established. The results illustrate that the robots generally developing the foundation of a stable behavior within a few dozen interactions, but beyond this the behavior is refined over time as the robot generalizes that distinction between properties 1 and 2 across the variations produced by properties 3 and 4. This refinement coincides with a gradual improvement in task performance after the first, relatively rapid phase of acquiring a generally successful strategy. The behavioral refinement over time reflects an individuation process in two separate habits (i.e., robot-diamond interactions and robot-circle interactions) becoming more distinct from one another to avoid interference between the two, e.g., suddenly switching to an established seeking behavior while in the process of avoiding because the dynamics of each resonate too similarly with a particular context. Although this is only a limited form of individuation—the distinction between the structures of the habits is already present, only their constitutive dynamics become more distinct—it nevertheless points to interesting developmental processes which occur even within this constrained model.

## 4.3. Limitations and Future Work

A criticism may be made of our investigation that the constraints and carefully arranged properties of the experiment mean that the model's internal adaptive mechanism serves an analogous function to an external optimization process such as an evolutionary algorithm. While this is the case here, because we are imposing a specific behavior on the system, the crucial difference is that our model would still have an adaptive and developmental gradient in the absence of such constructions. In the typical evolutionary approach a specific functional behavior is attractive in its own terms, *via* the fitness metric. However in our approach the particular functional behavior is made attractive through the relationship between the environmental dynamics and the internal processes of the robot and controller. Attractive behaviors will still arise for any specification of environmental dynamics and be meaningful in these terms.

Nevertheless, it is worth discussing the consequences of the network constraints. In particular, the robot's capacity to solve the task as we expect it to do so relies in part on the fact that it is incapable of assimilating the "wrong" environmental support into its sensorimotor structure. In other words, the dynamics associated with the diamond disappearing from the robot's view for much longer when it collides with the robot are always treated as disruptive regardless of context and of how many times it occurs, and *vice versa* for the circle. This high-level rigidity limits the ways in which the robot may adapt. In discussing the development of sensorimotor schemes in human development, Piaget discussed three different classes of adaptive processes of how instances of this disruption are resolved over time (Chapman, 1992; Boom, 2010): (1) the disruption is ignored without altering behavior; (2) The behavior alters to compensate disruptions which have previously been encountered, or (3) potential disruptions are anticipated and behavior is altered so that the disruption is not encountered at all. The second process would most accurately describe that which is occurring in the robots in this investigation, while the others

are not possible within the constraints that we have placed. The combination of these kinds of adaptive process is a key part of open-ended, autonomous development that is neicessary for sensorimotor agency. The obvious next step in terms of using the ASM-network to investigate sensorimotor agency is to remove the constraints at the network level, in a manner that allows precarious, self-maintaining structures to develop dynamically at that level.

Although the constraints we have placed on the model in this investigation limit the kinds of habits that may form autonomously, they allow for an analytically tractable investigation to demonstrate some of the model's capabilities. Our results provide a demonstration that the ASM-units are effective in producing behavior which supports the maintenance of a networked arrangement of such units that reflects a structure of sensorimotor coordinations. Furthermore, this alone is sufficient to produce a form of minimal cognitive behavior. However the model is also sufficiently generic that it is not necessary to have a pre-given network arrangement, engineered to align with a specific function, in order to produce coherent behavior. This opens the possibility to investigate self-organizing sensorimotor structures and adaptive autonomy in more depth in the future.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## AUTHOR CONTRIBUTIONS

FW and ME contributed to the writing of the manuscript. FW developed code and collected results. Both authors contributed to the article and approved the submitted version.

## FUNDING

FW's contribution was funded by the University of Auckland Doctoral Scholarship program. OAP fees were funded by the School of Computer Science, University of Auckland.

## REFERENCES

Barandiaran, X. E. (2017). Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. *Topoi* 36, 409–430. doi: 10.1007/s11245-016-9365-4

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* 17, 367–386. doi: 10.1177/1059712309343819

Barandiaran, X. E., and Di Paolo, E. A. (2014). A genealogical map of the concept of habit. *Front. Hum. Neurosci.* 8:522. doi: 10.3389/fnhum.2014.00522

Barandiaran, X. E., and Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artif. Life* 20, 5–28. doi: 10.1162/ARTL_a_00094

Bedia, M. G., Heras-Escribano, M., Cajal, D., Aguilera, M., and Barandiaran, X. E. (2019). "Towards modelling social habits: an organismically inspired evolutionary robotics approach," in *Artificial Life Conference Proceedings* (Newcastle), 341–348. doi: 10.1162/isal_a_00185.xml

Beer, R. D. (1996). "Toward the evolution of dynamical neural networks for minimally cognitive behaviour," in *From Animals to Animats*, eds P. Maes, M. J. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson (Cambridge, MA: MIT Press), 421–429.

Beer, R. D. (1997). The dynamics of adaptive behavior: a research program. *Robot. Auton. Syst.* 20, 257–289. doi: 10.1016/S0921-8890(96)00063-2

Beer, R. D. (2008). "The dynamics of brain-body-environment systems: a status report," in *Handbook of Cognitive Science*, eds P. Calvo and A. Gomila (San Diego, CA: Elsevier), 99–120. doi: 10.1016/B978-0-08-046616-3.00006-2

Boom, J. (2010). "Piaget on equilibration," in *The Cambridge Companion to Piaget*, eds U. Müller, J. I. M. Carpendale, and L. Smith (Cambridge: Cambridge University Press), 132–149. doi: 10.1017/CCOL9780521898584.006

Buhrmann, T., and Di Paolo, E. (2017). The sense of agency—a phenomenological consequence of enacting sensorimotor schemes. *Phenomenol. Cogn. Sci.* 16, 207–236. doi: 10.1007/s11097-015-9446-7

Buhrmann, T., Di Paolo, E. A., and Barandiaran, X. (2013). A dynamical systems account of sensorimotor contingencies. *Front. Psychol.* 4:285. doi: 10.3389/fpsyg.2013.00285

Chapman, M. (1992). "Equilibration and the dialectics of organization," in *Piaget's Theory: Prospects and Possibilities*, eds H. Beilin and P. B. Pufall (Hillsdale, NJ: Lawrence ErlbaumAssociates, Inc.), 39–59.

Di Paolo, E. (2019). Process and individuation: the development of sensorimotor agency. *Hum. Dev.* 63(Suppl. 3–4), 202–226. doi: 10.1159/000503827

Di Paolo, E., Buhrmann, T., and Barandiaran, X. (2017). *Sensorimotor Life: An Enactive Proposal.* New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780198786849.001.0001

Egbert, M. (2018). "Investigations of an adaptive and autonomous sensorimotor individual," in *The 2018 Conference on Artificial Life* (Tokyo: MIT Press), 343–350. doi: 10.1162/isal_a_00065

Egbert, M., and Cañamero, L. (2014). "Habit-based regulation of essential variables," in *Artificial Life 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, eds M. D. Egbert and X. E. Barandiaran (New York, NY), 168–175. doi: 10.7551/978-0-262-32621-6-ch029

Egbert, M. D., and Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Front. Hum. Neurosci.* 8:590. doi: 10.3389/fnhum.2014.00590

Froese, T., and Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artif. Intell.* 173, 466–500. doi: 10.1016/j.artint.2008.12.001

Georgeon, O. L., and Riegler, A. (2019). CASH only: constitutive autonomy through motorsensory self-programming. *Cogn. Syst. Res.* 58, 366–374. doi: 10.1016/j.cogsys.2019.08.006

Hutto, D. D., and Myin, E. (2012). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262018548.001.0001

Hutto, D. D., and Robertson, I. (2020). "Clarifying the character of habits: understanding what and how they explain," in *Habits, 1st Edn*, eds F. Caruana and I. Testa (New York, NY: Cambridge University Press), 204–222. doi: 10.1017/9781108682312.010

Iizuka, H., and Di Paolo, E. A. (2007). Toward spinozist robotics: exploring the minimal dynamics of behavioral preference. *Adapt. Behav.* 15, 359–376. doi: 10.1177/1059712307084687

Mirza, N. A., Nehaniv, C. L., and Dautenhahn, K. (2006). "Interaction histories: from experience to action and back again," in *Fifth International Conference on Development and Learning* (Bloomington, IN), 6.

Noe, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

O'Regan, J. K. (2011). *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness.* New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780199775224.001.0001

O'Regan, J. K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–973. doi: 10.1017/S0140525X01000115

Ramírez-Vizcaya, S., and Froese, T. (2019). The enactive approach to habits: new concepts for the cognitive science of bad habits and addiction. *Front. Psychol.* 10:301. doi: 10.3389/fpsyg.2019.00301

Ramírez-Vizcaya, S., and Froese, T. (2020). "Agents of habit: refining the artificial life route to artificial intelligence," in *The 2020 Conference on Artificial Life* (Montreal, QC: MIT Press), 78–86. doi: 10.1162/isal_a_00298

Vargas, P. A., Di Paolo, E. A., Harvey, I., and Husbands, P. (2014). "The horizons of evolutionary robotics," in *Intelligent Robotics and Autonomous Agents* (Cambridge, MA; London: The MIT Press). doi: 10.7551/mitpress/8493.001.0001

Woolford, F. M. G., and Egbert, M. D. (2020). "A precarious sensorimotor sequence reiterator for modelling enactive habits," in *The 2020 Conference on Artificial Life* (Montreal, QC: MIT Press), 771–779. doi: 10.1162/isal_a_00249

Woolford, F. M. G., and Egbert,M. D. (2019). Behavioural variety of a node-based sensorimotor-to-motor map. *Adapt. Behav.* 1–16. doi: 10.1177/1059712319839061

Zarco, M., and Egbert, M. D. (2019). "Different forms of random motor activity scaffold the formation of different habits in a simulated robot," in *Artificial Life Conference Proceedings* (Newcastle), 582–589.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Woolford and Egbert. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 01 June 2022
doi: 10.3389/fnbot.2022.848084

# Reach Space Analysis of Baseline Differential Extrinsic Plasticity Control

*Simon Birrell*, Arsen Abdulali and Fumiya Iida*

*Bio-Inspired Robotics Laboratory, Department of Engineering, University of Cambridge, Cambridge, United Kingdom*

The neuroplasticity rule Differential Extrinsic Plasticity (DEP) has been studied in the context of goal-free simulated agents, producing realistic-looking, environmentally-aware behaviors, but no successful control mechanism has yet been implemented for intentional behavior. The goal of this paper is to determine if "short-circuited DEP," a simpler, open-loop variant can generate desired trajectories in a robot arm. DEP dynamics, both transient and limit cycles are poorly understood. Experiments were performed to elucidate these dynamics and test the ability of a robot to leverage these dynamics for target reaching and circular motions.

**Keywords: Differential Extrinsic Plasticity, self-organization, robotic control, play, intrinsic motivation, neuroplasticity, reinforcement learning, complexity**

## 1. INTRODUCTION

Robot control is still very much a work in progress. While much has been learned of how humans and animals control their bodies (Winter, 2009), either outright or after a learning process, we still do not know enough to be able to design a robot that even approaches human dexterity. Classical control theory and more recently Reinforcement Learning (RL) have been extensively studied but are still subject to lack of robustness, the curse of dimensionality and unreasonably high learning times (Sutton and Barto, 2018).

One issue with these frameworks is the assumption that the brain directly controls the output of each available degree of freedom; typically a learning agent will adjust its body's motor torques at each time step to produce a desired result in a rigid body system within a given environment (see for example OpenAI Gym; Brockman et al., 2016). This is clearly not how biology tackles the problem. In a human, descending signals from the cortex pass through and are modified by interneurons with their own neuroplasticity mechanisms, which activate bundles of muscle fibers and drive an underactuated soft body with extremely complex dynamics (Pierrot-Deseilligny and Burke, 2005; Winter, 2009). On the face of it, we have no hope. If we can't reliably control a mathematically much simpler rigid robot, how could we possibly control an agent with a similar complexity to a human body?

On the other hand, could the complexity of the human body actually be a help and not a hindrance to the perception/control problem? The bio-inspired research agenda known as Embodied Intelligence suggests so (Pfeifer and Bongard, 2006; Cangelosi et al., 2015) and has spawned many different initiatives in this area. One approach is morphological computation, which investigates the ways that parts of the information processing burden can be offloaded to the body itself (Hauser et al., 2012; Müller and Hoffmann, 2017), through sensor morphology (as in the case of flies' eyes) (Iida and Nurzaman, 2016) or through the simplification of control (Eder et al., 2018). Physical reservoir computing uses the complexity of the body for general purpose computation

OPEN ACCESS

***Edited by:***
*Inês Hipólito,*
*Humboldt University of Berlin, Germany*

***Reviewed by:***
*Tianqi Wei,*
*University of Edinburgh, United Kingdom*
*Udaya B. Rongala,*
*Lund University, Sweden*

****Correspondence:***
*Simon Birrell*
*sab233@cam.ac.uk*

***Received:*** *03 January 2022*
***Accepted:*** *25 April 2022*
***Published:*** *01 June 2022*

***Citation:***
*Birrell S, Abdulali A and Iida F (2022) Reach Space Analysis of Baseline Differential Extrinsic Plasticity Control. Front. Neurorobot. 16:848084. doi: 10.3389/fnbot.2022.848084*

(Nakajima, 2020). The richness of behavior of the peripheral nervous system, providing fast-acting reflexes and hierarchical and coordinated control (Côté et al., 2018) has been less applied to robotic research, which almost exclusively models the control problem as the agent's brain directly driving motor torques. Finally, investigations into different neuroplasticity schemes show that a surprising variety of complex, environmentally-aware behaviors can be spontaneously generated from simple, biologically plausible neuroplasticity rules within sensorimotor loops (Zappacosta et al., 2018).

These latter neuroplasticity-generated spontaneous behaviors, detailed in the book "The Playful Machine" (Der and Martius, 2012) and in related papers (Der and Martius, 2015, 2017), can drive simulated agents to explore and react to their environments in a manner that is highly suggestive of natural behaviors without building in any goals or higher-level planning of any sort. The most recent iteration of this research uses a particular neuroplasticity scheme called Differential Extrinsic Plasticity (DEP) (Der and Martius, 2015; Pinneri and Martius, 2018) to generate intriguing behaviors that are tightly coupled with the environment: a four-legged creature will appear to search for and find ways to climb over a fence; a humanoid will eventually clamber out of a hole it is trapped in. From our external observer perspective, these embodied behaviors appear to be goal-driven, but yet they are not. DEP has emerged as an interesting and promising candidate plasticity rule, but to date no practical applications for it have yet been found. It is an autonomous goal-free controller rather than a useful control method or component in a larger system.

A complicating factor in the practical usage of DEP is the current lack of an analytical solution, despite the research time invested. In all likelihood, even simple DEP systems are too complex to be fully described analytically and so research has tended to be empirical, treating DEP as a pre-existing natural phenomenon. This is not an insurmountable issue; it places DEP within the context of related research into algorithmic information and complexity theory, both areas cited in theories of the development of the human brain (Hiesinger, 2021). The behavior of DEP may not be solvable analytically even if it is deterministic. It may be undecidable: the only way to determine the output being to run it in simulation.

Given this, how can we study DEP and map out its potential? First, we must simplify: by temporarily removing environmental feedback we can map out baseline behaviors for DEP, following the methods employed in Pinneri and Martius (2018). Second, we must test the control-ability and limits of what DEP can do: to what extent can higher order systems "request" particular behaviors, and how much coverage will these behaviors provide in the context of a given task?

This paper takes the first steps in this direction. By employing "short-circuit DEP" (see below) and with a simple test case, where the output of short-circuit DEP drives a simulated 2 degree of

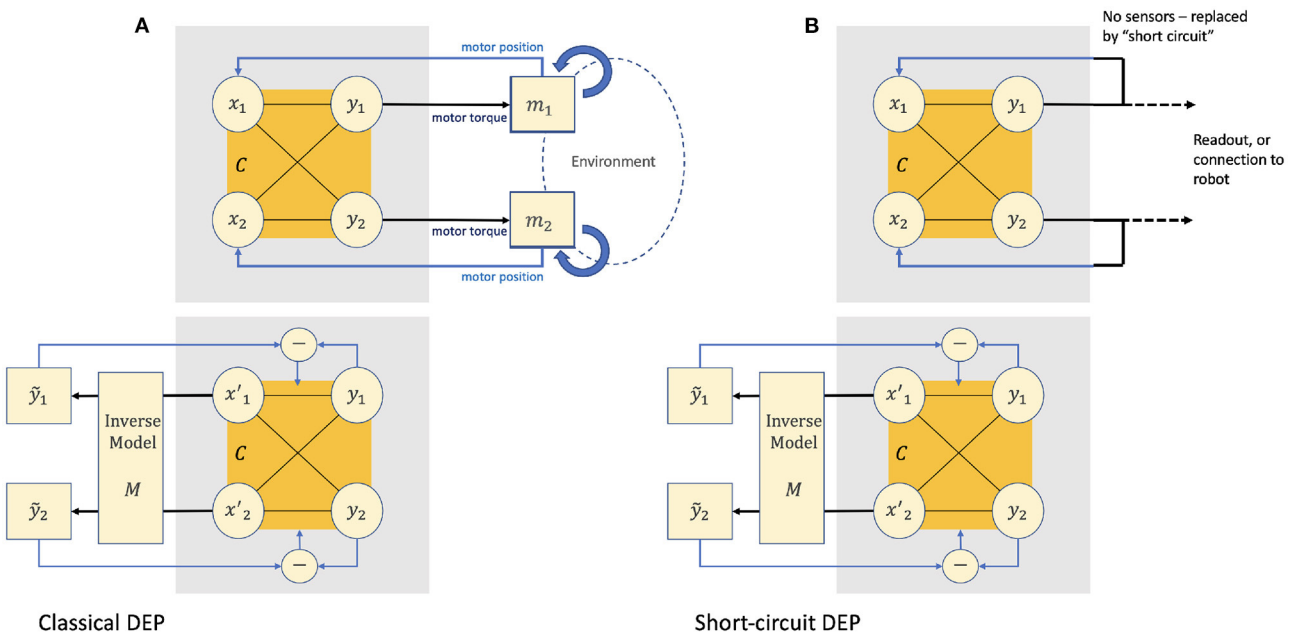



**FIGURE 1 |** In the "classical" version of Differential Extrinsic Plasticity (DEP) comprises two overlapping dynamical systems. [**(A)**, top] The input layer $x$ of a feed-forward neural network $C$ is driven by the motor positions. The output layer $y$ drives the motor torques. These motors operate on the agents body in a given environment (potentially with collisions) and the resulting motor positions $x^{t+1}$ will be fed back to the input to start the cycle again. [**(A)**, bottom] The positional information returned as $x^{t+1}$ is also fed to an inverse model that infers the rate of change of the motor torques $\tilde{y}$ that would have generated them in the absence of environmental feedback. The difference between $\tilde{y}$ and the actual rate of change $\dot{y}$ captures the environmental effects of the agent's actions. This difference is then used to modify the weights of C. In the "short-circuited version" of DEP [**(B)**, top and bottom] there are no motors or sensors; the output $y$ of the network is fed directly back to the input $x$.

freedom (DOF) robot arm, we show that DEP can be made to accomplish specific goals and that these goals cover a useful region of task space.

## 2. MATERIALS AND METHODS

### 2.1. How "Classical" Differential Extrinsic Plasticity Works

DEP describes a way of wiring motors and related sensors together with a neuroplasticity rule, such that a DEP-enabled agent produces a large set of "natural looking" behaviors that respond to interactions with the environment. Summarizing (Der and Martius, 2015; Pinneri and Martius, 2018), this section describes the equations that define the thermoplasticity rule for the "classic" version of DEP (see **Figure 1A**).

For a two-layer artificial neural network with input layer $x_i$, output layer $y_i$, weights $C_{ij}$, biases $h_i$, and a *tanh* activation function, the output activation is given by:

$$y_i = tanh(\sum_{j=1}^{n} C_{ij}x_j + h_i) \quad (1)$$

A simple feedback controller for an agent with rotary motors may then be constructed where $y_i$ drives motor torques and $x_i$ is driven by the resulting motor positions (see **Figure 1A** for a 2 degree of freedom example). In itself, this is not a very interesting controller, although given that the motor positions are ultimately determined not only by the applied torques but also by the body in which they're embedded and its interaction with the environment, nor is it trivial. This neural controller, the body and the environment together form a single dynamical system.

The behavior of this dynamical system can be overlaid by a second dynamical system driven by neural plasticity, that is, the evolution over time of the controller's weights. Many plasticity schemes have been studied (see **Table 1**). Hebbian learning modifies the weights based on the product of pre and post-synaptic activations[1]. Differential Hebbian Learning is similar (Zappacosta et al., 2018), but uses the product of the rates of change of the two activations.

Differential Extrinsic Plasticity extends Differential Hebbian learning by introducing an inverse model $F$ that maps the rate of

**TABLE 1 |** Different plasticity schemes.

| Plasticity scheme | Update rule |
|---|---|
| Hebbian learning | $\tau\dot{C}_{ij} = x_i y_j$ |
| Differential Hebbian learning | $\tau\dot{C}_{ij} = \dot{x}_i \dot{y}_j$ |
| Differential extrinsic plasticity | $\tau\dot{C}_{ij} = \dot{x}_i(\dot{y}_j + \delta\dot{y}_j) - C_{ij}$ |

[1]Using Hebbian learning here would give a system that resembles a continuous variable Hopfield Network, but with normalization and an inverse model.

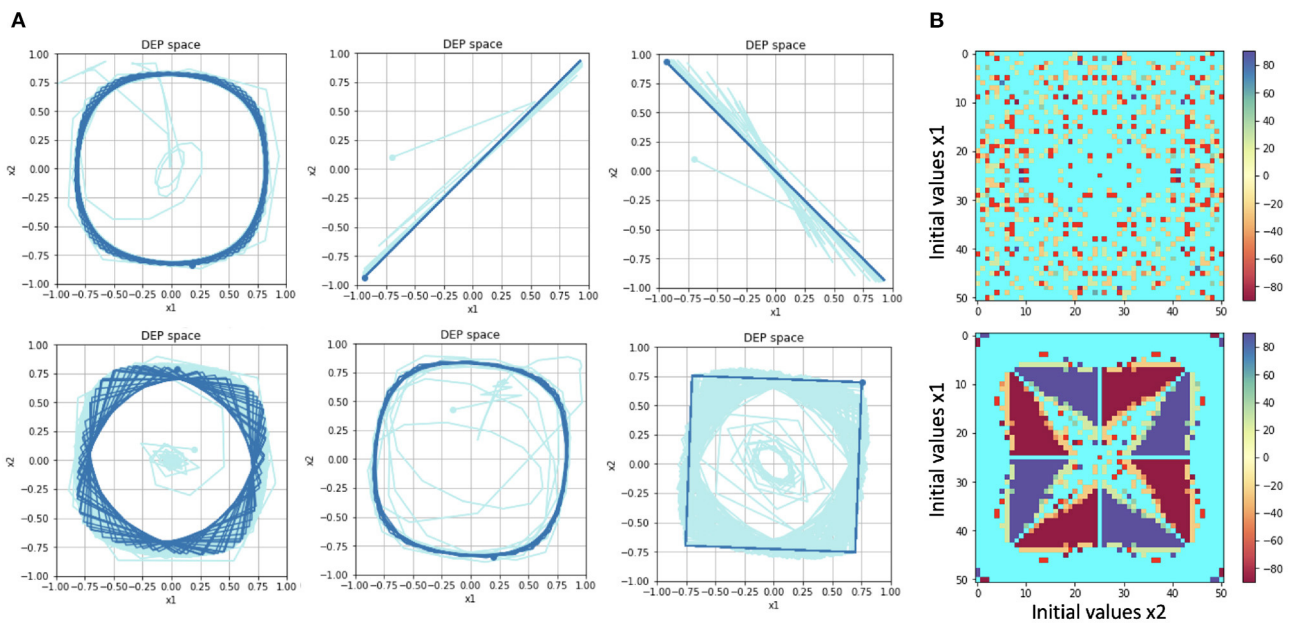



**FIGURE 2 | (A)** Examples of limit cycles reached by "Short-circuit" DEP. The phase diagrams show the trajectory of the system along the dimensions $x_1$ and $x_2$. The second and third trajectories oscillate between two endpoints. The other trajectories are all rotational. **(B)** Maps of the attractors reached based on initial values of $x_1, x_2$. The color bar refers to the rotational angle of the attractor, in the case of rotational attractors. Cyan refers to the non-rotational attractors shown in the second and third examples in **(A)**. The resulting map is shown in the top row. The map in the lower row reproduces the one shown in Pinneri and Martius (2018), but to generate it requires slightly altering the DEP algorithm (see text). Our version lacks their basins of attraction.

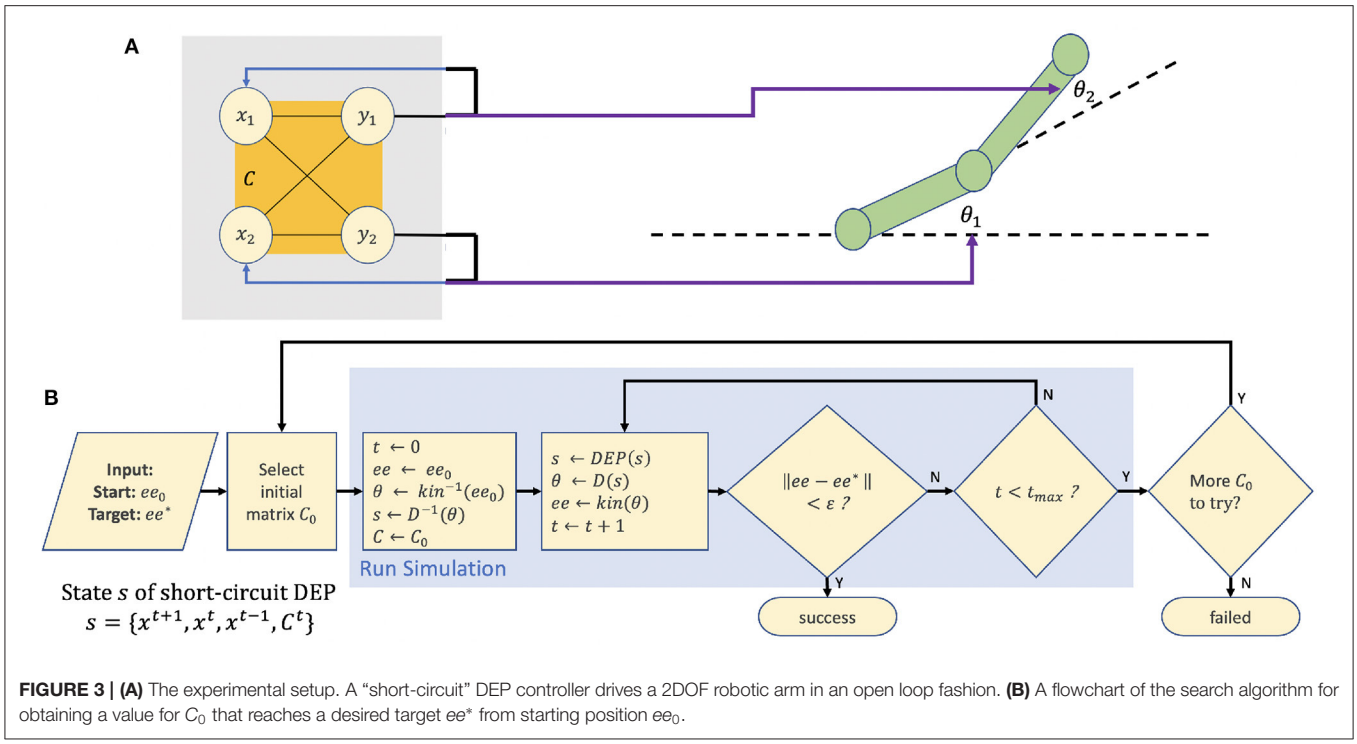



**FIGURE 3 | (A)** The experimental setup. A "short-circuit" DEP controller drives a 2DOF robotic arm in an open loop fashion. **(B)** A flowchart of the search algorithm for obtaining a value for $C_0$ that reaches a desired target $ee^*$ from starting position $ee_0$.

change of received sensor values $\dot{x}^{t+1}$ back to the inferred rate of change of motor torques $\tilde{\dot{y}}^t$ that caused them:

$$\tilde{\dot{y}}^t = F(\dot{x}^{t+1}) \tag{2}$$

In most DEP implementations the inverse model $F$ is implemented as a simple matrix $M$ such that

$$\tilde{\dot{y}}^t = M\dot{x}^{t+1} \tag{3}$$

and, as in this paper, it is often assumed to be the identity matrix. $F$ isn't required to be strictly accurate to reproduce DEP's behavior (Der and Martius, 2015).

The revised update rule uses $\tilde{\dot{y}}$ in place of $\dot{y}$ and adds a damping term. Dropping the time superscript $t$:

$$\tau \dot{C}_{ij} = \dot{x}_i \tilde{\dot{y}}_j - C_{ij} \tag{4}$$

One way to think about $\tilde{\dot{y}}$ is as the sum of the real historical value for $\dot{y}$ at $t$ plus an error term $\delta\dot{y}$ with respect to the model $F$.

$$\tilde{\dot{y}} = \dot{y} + \delta\dot{y} \tag{5}$$

This substitution is shown in the final row of **Table 1**. Comparing it to the scheme for Differential Hebbian Learning shows how this "unexpected" environmental feedback is incorporated into the weight updates.

The weight matrix $C$ is normalized to $\hat{C}$ at each time step with a factor $\kappa$ and a parameter $\rho$ that prevents a division by zero.

$$\hat{C} \leftarrow \kappa C/(\|C\| + \rho) \tag{6}$$

Finally, the activation rule is modified from Equation (1) to use the normalized $\hat{C}$ rather than $C$:

$$y_i = tanh(\sum_{j=1}^{n} \hat{C}_{ij} x_j + h_i) \tag{7}$$

The combination of these two overlaid dynamical systems produces an agent that cycles through a series of complex behaviors that are responsive to environmental feedback.

## 2.2. How "Short-Circuit" DEP Works

A simplified version of DEP was used in Pinneri and Martius (2018) for an empirical analysis of its behaviors. In this configuration there are no motors or sensors; the system output $y$ is connected directly back to the system input $x$ (**Figure 1B**). As $\tilde{\dot{y}}^t = M\dot{x}^{t+1}$ and $\dot{x}^{t+1} = \dot{y}^t$ and $M$ is the identity matrix the update rule simplifies to

$$\tau \dot{C}_{ij} = \dot{x}_i \dot{y}_j - C_{ij} \tag{8}$$

This is effectively Differential Hebbian Learning with damping and normalization.

By eliminating the environment, the behaviors generated can be simplified to a set of predictable limit cycles, examples of which are shown in **Figure 2A**. The limit cycles reached depend on the initial conditions of the system, in particular the initial values for $x_1, x_2, y_1, y_2, \dot{x_1}, \dot{x_2}, \dot{y_1}, \dot{y_2}$, and $C$. In Pinneri and Martius (2018), all initial values were held constant except for $x_1, x_2$.

Following that paper, a map of the attractors reached based on differing initial conditions for $x_1, x_2$ is shown in **Figure 2B**. For each fixed point, the final 2x2 $\hat{C}$ matrix, is considered to be a

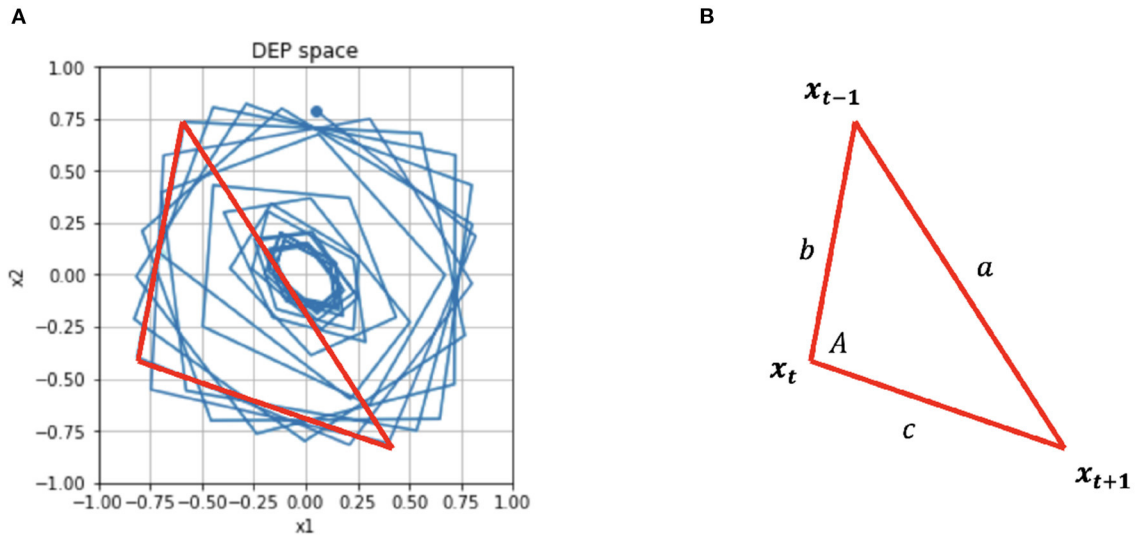



**FIGURE 4 | (A)** For the circular controller, $\theta_2$ is derived from the angle between subsequent vertices of $x$. **(B)** This angle A is calculated by simple geometry.

rotational matrix and the corresponding angle is assigned a color. There are two cases of non-rotational matrices: a zero matrix (which is assigned bright red) and period-2 oscillations, such as the second and third examples in **Figure 2A**, which are assigned cyan. The resulting map is shown in **Figure 2B** (top).

It should be noted that (Pinneri and Martius, 2018) obtained a different pattern, as that paper used code that inadvertently reset the $C$ matrix to zero at $t = 2$, generating different dynamics (private communication). Their results were reproduced (with the necessary code modification) in **Figure 2B** (bottom). For our experiments we followed the strict interpretation of the DEP equations. The attractors identified are the same, but the attractor map with respect to initial conditions is different; the "basins of attraction" cited in that paper being absent. In our opinion, these basins are an artifact of the previous code base and not intrinsic to DEP as such.

In the present paper's experiments, as well as $x_1, x_2$, the initial value $C_0$ of the matrix $C$ is also varied. It was discovered that choosing different values for $C_0$ elicits different trajectories and ultimate limit cycles for each combination of the initial values $x_1, x_2$. One way of looking at this is to say that different $C_0$ can select different behaviors for a given initial $x_1, x_2$.

## 2.3. The Experimental Setup

In the two experiments described, the "short circuit" DEP system is used to drive a simple 2 degree of freedom robotic arm (see **Figure 3A**).

The state $s$ of the short-circuit DEP system can be fully captured as

$$s = \{x^{t+1}, x^t, x^{t-1}, C^t\} \tag{9}$$

so that at each timestep $s^{t+1} \leftarrow DEP(s^t)$.

We can then use a robot arm with segment lengths $l_1, l_2$, here 0.5 m, to "read out" the state $s$ of DEP. The joint angles $\theta$, comprising $\theta_1, \theta_2$, are driven by a "driver" function $D$ that is specific to a given task type, such that

$$\theta = D(s) \tag{10}$$

Note that this is an open-loop controller. None of the reported benefits of environmentally-aware "Classic" DEP are used here, in line with the goal of learning to control a very simple DEP system. The position of the robot's end effector can be considered as a simple transformation or readout of DEP's internal state $s$.

Two types of task are considered. In the first, the goal is for the robot arm's end effector that starts at position $ee_0$ to **reach** an arbitrary target position $ee^\star$. For this type of task, function $D(s) = D_{reach}(s)$ is simply

$$\begin{aligned} \theta &= D_{reach}(\{x^{t+1}, x^t, x^{t-1}, C^t\}) \\ &= \pi x^{t+1} \\ &= \pi y \end{aligned} \tag{11}$$

In other words, the output of $y$ of short-circuit DEP directly drives the motor angles $\theta$.

In the second type of task, the goal is for the end effector to trace a **circular** trajectory of arbitrary radius $r$. Here, $D(s) = D_{circle}(s)$ and we leverage the angle $A$ between the vectors $x^{t+1} - x^t$ and $x^t - x^{t-1}$. See **Figure 4** for the simple geometry that defines $a, b, c$. Then, the two joint angles $\theta_1, \theta_2$ can be defined in the new driver function:

$$\begin{aligned} \theta &= D_{circular}(\{x^{t+1}, x^t, x^{t-1}, C^t\}) \\ \theta_i &= \begin{cases} \omega t & \text{if } i = 1 \\ cos^{-1}\left(\frac{b^2+c^2-a^2}{2bc}\right) & \text{if } i = 2 \end{cases} \end{aligned} \tag{12}$$

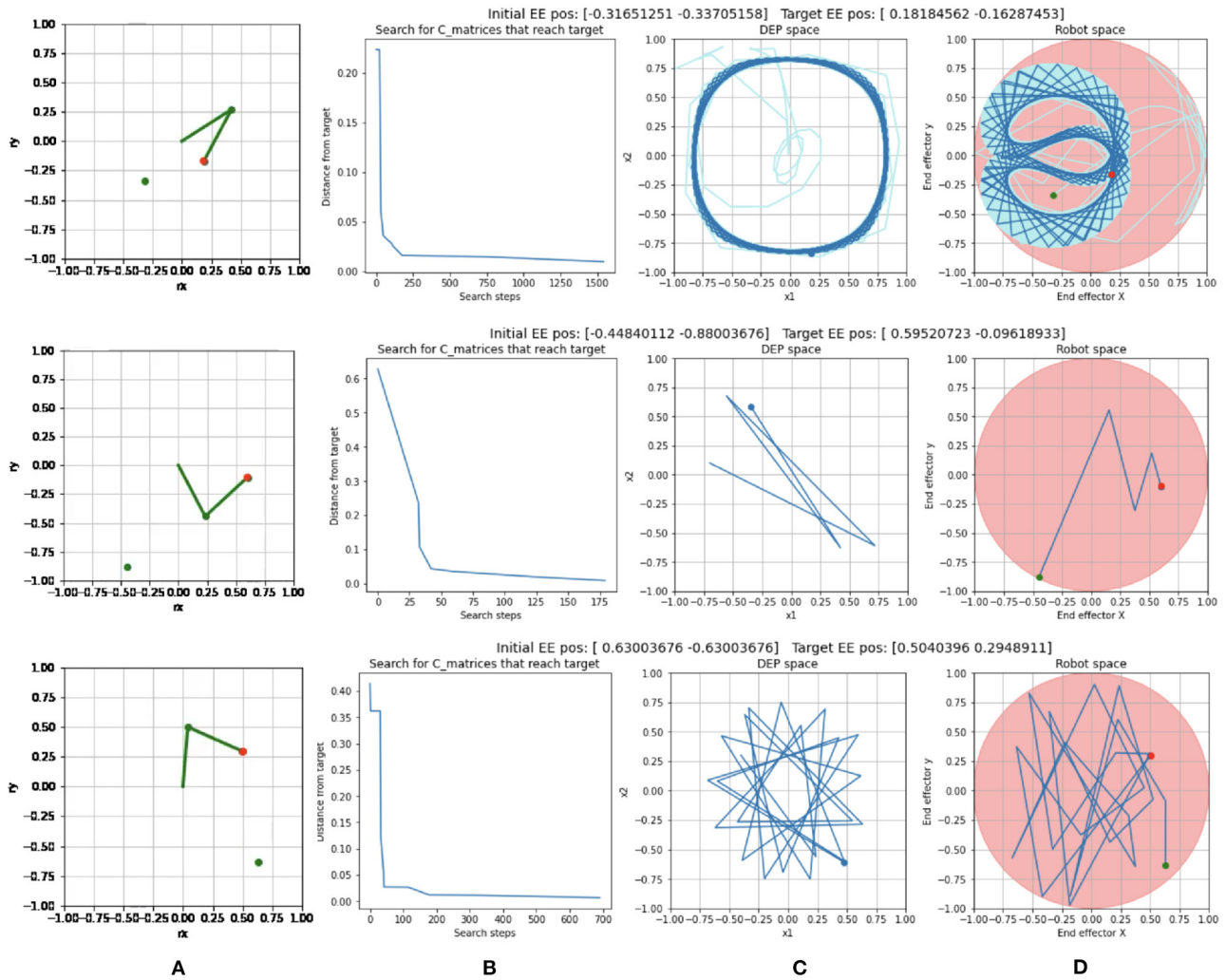



**FIGURE 5 |** Examples of Reaching trials with: **(A)** different end effector start positions (green dot) and target positions (red dot). **(B)** The progress of the search for a solution C matrix. **(C)** The solution trajectory in DEP space for the successful trial. **(D)** The solution trajectory in robot space for the successful trial. The most recent positions are shown in dark blue while the early part of the trajectory is in light blue.

At a fixed point of $C$, $\hat{C}^{t+1} \rightarrow C^t$ and if $|x| << 1$,

$$x^{t+1} = tanh(\hat{C}x^t), \quad x^{t+1} \approx \hat{C}x^t \tag{13}$$

Under these conditions, $x$ is rotating around the origin in DEP space and the angle between every other point is approximately constant. As this angle drives $\theta_2$ then $\theta_2$ will also be a constant. $\theta_1$ is a constant, so the robot will describe a circle.

## 2.4. The Search Algorithm for $C_0$

Given an input of an initial end effector position $ee_0$ and target position or trajectory $ee^\star$, our goal is to obtain an initial matrix $C_0$ that will drive the system to reach $ee^\star$. $C_0$ is obtained by a search algorithm, detailed in **Figure 3B**.

The $2 \times 2$ matrix $C_0$ has four parameters that here each vary between $-1$ and $+1$. The algorithm linearly divides the range of each parameter into eight values, giving $8 \times 8 \times 8 \times 8 = 4{,}096$ possible values for $C_0$. A simple grid search is performed, with each value being trialed in a rollout of 20,000 time steps.

In the case of the reaching task, at each time step of the rollout, if the distance between $ee_t$ and $ee^\star$ is within a given tolerance $\epsilon$, then success is declared. 10 random starting positions $ee_0$ and 10 random targets $ee^\star$ were combined to give 100 trials, each of which is an execution of the algorithm in **Figure 3B**.

In the case of the circular task, success is declared after a full rotation of the end effector, where the mean squared radius error with respect to $r$ is less than $\epsilon$. Five random starting positions $ee_0$ and five random radii $r^\star$ were combined to give 25 trials, each of which is an execution of the algorithm in **Figure 3B**.

The experiments were implemented in Python on Jupyter notebooks. The full source code may be downloaded from GitHub[2], inspected and run.

[2]https://github.com/SimonBirrell/dep-control

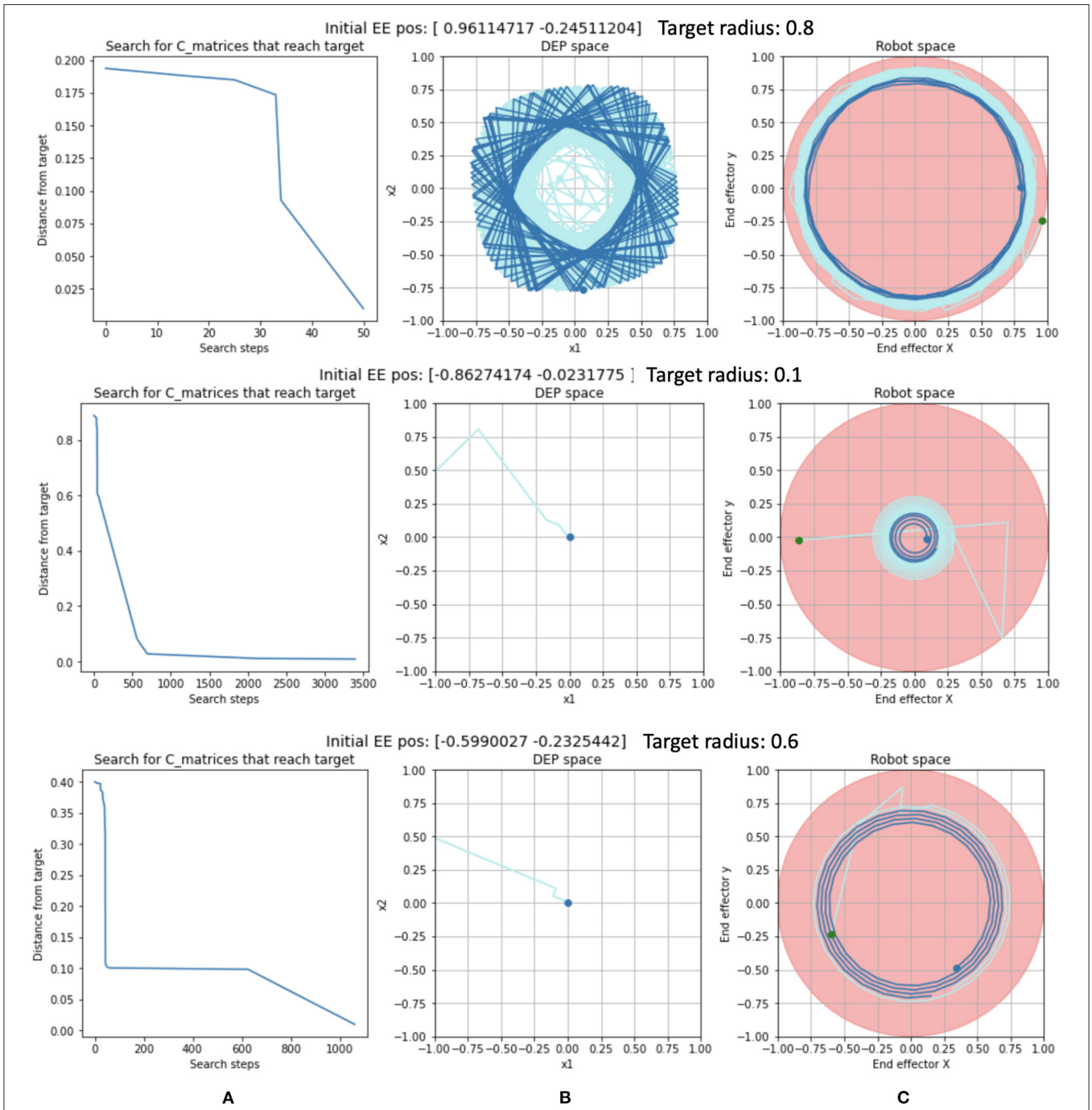



**FIGURE 6 |** Examples of Circular Trajectory trials with: **(A)** The progress of the search for a solution C matrix. **(B)** The solution trajectory in DEP space. **(C)** The solution trajectory in robot space. The most recent positions are shown in dark blue while the early part of the trajectory is in light blue.

## 3. RESULTS

The trajectories in DEP space produced in the experiments generally consisted of a transient phase where the system "wanders" in $x_1, x_2$ followed by a limit cycle phase. The Reaching task leveraged both transient and limit cycle phase, while the Circular task leveraged the limit cycles.

### 3.1. The Reaching Task

One hundred trials of the Reaching task were performed. In every case, the system reported success: it found a path to all end effector targets from all end effector starting positions. The tolerance $\epsilon|$ had a value of 0.01 m.

Trajectory examples are show in **Figure 5**. In the example in the top row, the search algorithm tested 1,541 $C_0$ matrices

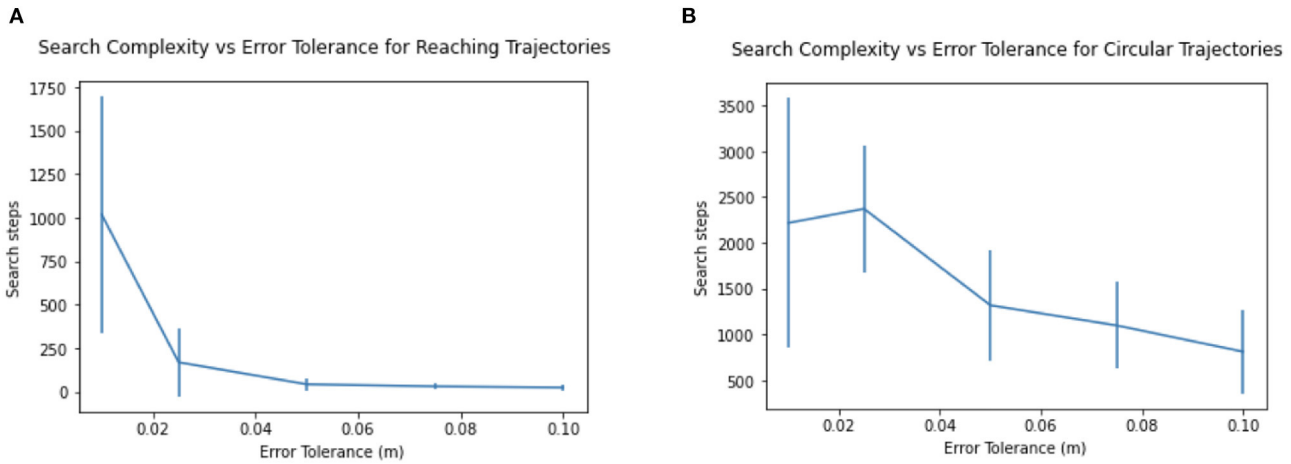



**FIGURE 7 |** Search complexity and variance increases with lower error tolerance, for **(A)** Reaching trajectories and **(B)** Circular trajectories.

(**Figure 5B**) before finding a value that caused the end effector (**Figure 5A**) to reach the desired target. The solution itself in DEP space (**Figure 5C**) and robot space (**Figure 5D**) show that the system had entered a rotational limit cycle before reaching the target.

A second example, in the second row of **Figure 5** shows a contrasting example where a solution was found after testing only 179 search steps. In the solution, the system was still in a transient phase when it hit the target, at only six time steps into the rollout.

## 3.2. The Circular Trajectory Task

Twenty-five trials of the Circular task were performed. In every case, the system reported success: it managed to describe a circular trajectory of at least one rotation where the mean squared radius error with respect to the desired radius was less than $\epsilon$, in this case 0.01 m.

Trajectory examples are shown in **Figure 6**. In the example in the top row, the search algorithm required a mere 50 steps (**Figure 6A**) to find a limit cycle in DEP space (**Figure 6B**) that completed a circle of the desired radius (**Figure 6C**) after 551 time steps of the rollout.

In the second example, in the second row of **Figure 6**, the search algorithm required 3,412 search steps (out of a maximum of 4,096) (**Figure 6A**) to find a solution in DEP space (**Figure 6B**) that completed a circle after 332 time steps of the rollout.

The relationship of search time to tolerance $\epsilon$ can be seen in **Figure 7**. For lower, more stringent, error tolerances $\epsilon$, the number of search steps required increases, as does its variance. Increasing the tolerance required for reaching even slightly (say from 0.01 to 0.025 m) reduces the search steps required by 75%.

# 4. DISCUSSION AND FUTURE WORK

The controller described in this paper is unlikely to signal the end of inverse kinematics. To borrow Dr. Johnson's phrase, it "is like a dog's walking on his hinder legs. It is not done well; but you are surprised to find it done at all" (Boswell, 1791). Why do these results, and DEP in general, matter? We can answer in three ways.

## 4.1. DEP as a Control Mechanism

First, what is the prognosis for DEP as a control system? The present controller has reduced a high dimensional control problem to one of simple selection of one of 4,096 different discrete values of the $C_0$ matrix. The original motivation for this paper was to find a way to leverage DEP within the context of Reinforcement Learning. $C_0$ provides a low dimensional interface for higher level systems to exploit. Yet most of the solutions are indirect, taking time for the end effector to reach its goal.

The search algorithm could be extended to optimize for lower time steps to reach the desired target position or trajectory. Different trajectory types could be produced with different driving functions, although fewer functions would be preferable to more. Driving functions could be abstract, as they are here, or derived from physical models of body elements, such as springs, tissue, or muscles.

There is scope for improving the search algorithm itself from a simple grid search, depending on what patterns, if any, can be found in the mapping of target to $C_0$. Are there basins of attraction for $C_0$? Is this controller learnable in a way that generalizes?

Once understanding of the core behavior of "short-circuit" DEP has improved, environmental awareness, one of the core supposed advantages of the neuroplasticity rule, could be reintroduced. This opens the way to recovery from perturbations and short term, "reflex" reactions to changes in the environment.

## 4.2. The Study of DEP

A continuing expressed frustration in the DEP literature is the lack of a full analytical treatment of DEP behavior. That may be due a lack of human resources applied to the problem, or it may be that a full treatment is simply intractable. Some algorithms are mathematically "undecidable," which is to say that their

behavior cannot be predicted without executing the algorithm itself. Perhaps DEP falls into this category.

In either case, this paper follows recent work in taking an empirical, engineering approach to analysing DEP, rather than a theoretical treatment. There remain many questions to be answered.

DEP has produced some fascinating simulations, with realistic looking and intriguing behaviors, such as gait switching, overcoming obstacles, and interaction with devices such as handles. How much of the observed behaviors are due to DEP as a neuroplasticity rule and how much are due to the particular body morphology of the simulated agents? Passive walkers also produce realistic behaviors and respond to the environment in a limited way, yet they have no neuroplasticity at all. Clearly, the agents behavior is generated by the complete system of neuroplasticity plus body plus environment. How we can disentangle the contributions of each?

Finally, does DEP scale? What are the limit cycles of higher dimensional DEP systems? Our understanding of DEP behavior is only just beginning.

### 4.3. Leveraging Pre-existing Complexity

DEP is an example of self-organization in action: of complexity generated from simple rules. Self-organization is easy to spot, but hard to design, yet may be necessary to enable long-term learning processes such as evolution to work effectively (Kauffman, 1995). Classical DEP is a system that, in that evocative phrase, exists "on the edge of chaos," producing a rich set of behaviors even in the "short circuit" version. Is this complexity useful to agents, or is it a simple artifact?

The leverage of pre-existing complex behaviors is seen in Physical Reservoir Computing (PRC), a field that applies a thin layer of learning over highly complex, pre-existing dynamics in a real or simulated body. A PRC system leverages a set of dynamical behaviors as if they were basis functions and combines them using a shallow artificial neural network. The network can then be trained to perform some desired function. The dimensionality of a problem that might require training a very deep neural network has been reduced to that of training a shallow one.

In the case of PRC, the pre-existing complexity is physical. In other cases it may be algorithmical. A curious example is the history of procedural content generation in computer games (Smith, 2015). The practice originated over 40 years ago with the need to generate details of thousands of planets in highly resource-constrained computers. Rather than store such details, they were generated from the Fibonacci sequence, passed through an interpretive function analogous to our "driver function." By using a predictable mathematical sequence that has inherent complexity, a vast amount of content could be generated *ex nihilio*. Other examples of Algorithmic Information have been studied, such as the "undecidability" and Turing completeness of Rule 110 (Cook et al., 2004).

What is unclear is whether and to what extent nature has leveraged these potential sources of complexity. In developmental biology, there is a gap between the information specified in the genome and the complexity of the end product (Hiesinger, 2021). In learning there is a gap between the mechanisms we have available and the complexity of the problems to solve. Does pre-existing complexity play a part in closing this gap? Is DEP an example of this?

Differential Extrinsic Plasticity remains a fascinating phenomenon. Neuroplasticity remains an under-explored component of Embodied Intelligence and a rich opportunity for future work.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: GitHub, https://github.com/SimonBirrell/dep-control.

## AUTHOR CONTRIBUTIONS

SB researched DEP, performed the experiments, and wrote the paper under the supervision of FI, with suggestions and comments from AA. All authors contributed to the article and approved the submitted version.

## FUNDING

This project was possible thanks to EPSRC Grant EP/L01-5889/1, the Royal Society ERA Foundation Translation Award (TA160113), EPSRC Doctoral Training Program ICASE Award RG84492 (cofunded by G's Growers), EPSRC Small Partnership Award RG86264 (in collaboration with G's Growers), and the BBSRC Small Partnership Grant RG81275.

## ACKNOWLEDGMENTS

The authors would like to thank Josie Hughes for her help and encouragement. Also to Georg Martius and Cristina Pinneri for giving access to the code behind their paper.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2022.848084/full#supplementary-material

**Supplementary Video 1 |** Evolution of the dynamics in the first row of **Figure 5**.

**Supplementary Video 2 |** Evolution of the dynamics in the second row of **Figure 5**.

**Supplementary Video 3 |** Evolution of the dynamics in the third row of **Figure 5**.

**Supplementary Video 4 |** Evolution of the dynamics in the first row of **Figure 6**.

**Supplementary Video 5 |** Evolution of the dynamics in the second row of **Figure 6**.

**Supplementary Video 6 |** Evolution of the dynamics in the third row of **Figure 6**.

## REFERENCES


Boswell, J. (1791). *The life of Samuel Johnson, LL.D*. London: Charles Dilly; Henry Baldwin.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., et al. (2016). Openai gym. *arXiv [Preprint] arXiv:1606.01540*. doi: 10.48550/arXiv.1606.01540

Cangelosi, A., Bongard, J., Fischer, M. H., and Nolfi, S. (2015). "Embodied intelligence," in *Springer Handbook of Computational Intelligence*, eds J. Kacprzyk and W. Pedrycz (Springer), 697–714. doi: 10.1007/978-3-662-43505-2_37

Cook, M., et al. (2004). Universality in elementary cellular automata. *Complex Syst*. 15, 1–40.

Côté, M.-P., Murray, L. M., and Knikou, M. (2018). Spinal control of locomotion: individual neurons, their circuits and functions. *Front. Physiol*. 9:784. doi: 10.3389/fphys.2018.00784

Der, R., and Martius, G. (2012). *The Playful Machine - Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Berlin Heidelberg: Springer.

Der, R., and Martius, G. (2015). Novel plasticity rule can explain the development of sensorimotor intelligence. *Proc. Natl. Acad. Sci. U.S.A*. 112, E6224–E6232. doi: 10.1073/pnas.1508400112

Der, R., and Martius, G. (2017). Self-organized behavior generation for musculoskeletal robots. *Front. Neurorobot*. 11:8. doi: 10.3389/fnbot.2017.00008

Eder, M., Hisch, F., and Hauser, H. (2018). Morphological computation-based control of a modular, pneumatically driven, soft robotic arm. *Adv. Robot*. 32, 375–385. doi: 10.1080/01691864.2017.1402703

Hauser, H., Ijspeert, A., Füchslin, R., Pfeifer, R., and Maass, W. (2012). The role of feedback in morphological computation with compliant bodies. *Biol. Cybern*. 106, 595–613. doi: 10.1007/s00422-012-0516-4

Hiesinger, P. R. (2021). *The Self-Assembling Brain: How Neural Networks Grow Smarter/Peter Robin Hiesinger*. Princeton, NJ: Princeton University Press. doi: 10.1515/9780691215518

Iida, F., and Nurzaman, S. G. (2016). Adaptation of sensor morphology: an integrative view of perception from biologically inspired robotics perspective. *Interface Focus* 6:20160016. doi: 10.1098/rsfs.2016.0016

Kauffman, S. A. (1995). *At Home in the Universe : The Search for Laws of Self-Organization and Complexity/Stuart Kauffman*. New York, NY; Oxford: Oxford University Press.

Müller, V. C., and Hoffmann, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artif. Life* 23, 1–24. doi: 10.1162/ARTL_a_00219

Nakajima, K. (2020). Physical reservoir computing—an introductory perspective. *Japanese J. Appl. Phys*. 59:060501. doi: 10.35848/1347-4065/ab8d4f

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3585.001.0001

Pierrot-Deseilligny, E., and Burke, D. (2005). *The Circuitry of the Human Spinal Cord: Its Role in Motor Control and Movement Disorders*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511545047

Pinneri, C., and Martius, G. (2018). "Systematic self-exploration of behaviors for robots in a dynamical systems framework," in *Proceedings of Artificial Life XI* (Cambridge, MA: MIT Press), 319–326. doi: 10.1162/isal_a_00062

Smith, G. (2015). "An analog history of procedural content generation," in *FDG* (Boston, MA). doi: 10.1201/9781315313412-9

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction, 2nd Edn*. Cambridge, MA: The MIT Press.

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470549148

Zappacosta, S., Mannella, F., Mirolli, M., and Baldassarre, G. (2018). General differential Hebbian learning: capturing temporal relations between events in neural networks and the brain. *PLoS Comput. Biol*. 14:e1006227. doi: 10.1371/journal.pcbi.1006227


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Birrell, Abdulali and Iida. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

HYPOTHESIS AND THEORY
published: 14 June 2022
doi: 10.3389/fnbot.2022.815850

# Carving Up Participation: Sense-Making and Sociomorphing for Artificial Minds

*Robin L. Zebrowski*[1*†] *and Eli B. McGraw*[2†]

[1] *Departments of Philosophy, Psychology, and Computer Science, Cognitive Science Program, Beloit College, Beloit, WI, United States,* [2] *Independent Scholar, Berryville, VA, United States*

AI (broadly speaking) as a discipline and practice has tended to misconstrue social cognition by failing to properly appreciate the role and structure of the interaction itself. Participatory Sense-Making (PSM) offers a new level of description in understanding the potential role of (particularly robotics-based) AGI in a social interaction process. Where it falls short in distinguishing genuine living sense-makers from potentially cognitive artificial systems, sociomorphing allows for gradations in how these potential systems are defined and incorporated into asymmetrical sociality. By side-stepping problems of anthropomorphism and muddy language around it, sociomorphing offers a framework and ontology that can help researchers make finer distinctions while studying social cognition through enactive sociality, PSM. We show here how PSM and sociomorphing, taken together and reconceived for more than just social robotics, can offer a robust framework for AGI robotics-based approaches.

**Keywords: anthropomorphism, sociomorphing, artificial general intelligence, Participatory Sense-Making, enactivism, social cognition, social robotics**

OPEN ACCESS

***Edited by:***
*Inês Hipólito, Humboldt University of Berlin, Germany*

***Reviewed by:***
*Massimiliano Lorenzo Cappuccio, United Arab Emirates University, United Arab Emirates*
*Antonio Chella, University of Palermo, Italy*
*Elena Clare Cuffari, Franklin & Marshall College, United States*
*Ekaterina Sangati, Okinawa Institute of Science and Technology Graduate University, Japan*

***Correspondence:**
*Robin L. Zebrowski*
*zebrowsr@beloit.edu*

*†These authors have contributed equally to this work and share first authorship*

***Received:*** *15 November 2021*
***Accepted:*** *24 May 2022*
***Published:*** *14 June 2022*

***Citation:***
*Zebrowski RL and McGraw EB (2022) Carving Up Participation: Sense-Making and Sociomorphing for Artificial Minds. Front. Neurorobot. 16:815850. doi: 10.3389/fnbot.2022.815850*

## INTRODUCTION

Seibt et al. (2020a) argue that social robotics/human robot interaction has a "description problem" insofar as it lacks a multidisciplinary set of terminology for describing apparently-social interactions between (at least) people and social robots. They point out that some capacities in these interactions can literally, rather than figuratively, be ascribed to robots, but that our current ontologies for making sense of these interactions fail us. Sociomorphing is the direct perception of real social capacities in agents or systems, including non-human agents. They say, "Such interactions, we proposed, should not by default and in all cases be viewed as involving a mental operation where *fictional human capacities for social interaction are imaginatively projected* onto the robot; rather, heeding a suitably wide understanding of sociality, we should allow for human social behavior towards robots to be guided by *direct (and possibly implicit) perceptions of actual non-human capacities for social interaction*" (2020, p. 63, emphasis in original). While Seibt's work here focuses on social robotics in particular, the terminology and conceptual ontology may be applicable outside of just social robotics work, and might offer important insights into artificial general intelligence (AGI)[1] work as it relates to robotics more broadly (Seibt, 2017). We propose to tease apart the conceptual framework offered by Seibt et al. that combines sociomorphing with

[1] We use AGI here in the more traditional sense of the term, as it overlaps with concepts of machine consciousness, machine mindedness, and early understandings of AI in general as an instantiation not merely of human-level intelligence, but of the concomitant mind as what is doing the thinking (Newell and Simon, 1961; Fuchs, 2021).

their larger project that then applies this new concept to an ontology of simulation. Since our project here aims at showing how this framework could help AGI work, rather than robotics work that focuses on more surface-level social interactions, we abandon Seibt et al.'s ontological commitments to simulation and imitation in favor of using these concepts in an entirely different kind of project, that of robotics-based AGI work. To this end, we include an important discussion of enactive social cognition, without which the sociomorphing work cannot get a foothold for AGI. Participatory Sense-Making (PSM) (De Jaegher and Di Paolo, 2007) offers a new level of enactive description with regard to interaction among multiple autonomous agents. As argued elsewhere (Zebrowski and McGraw, 2021) PSM is a missing and valuable variable in robotic systems as they relate to AGI (see text footnote 1). In attempting to apply PSM to robotic systems, though, it appears mysterious how and when we might attribute certain capacities, such as sentience or autonomy, to artificial systems. However, here, we propose that Seibt et al.'s notion of sociomorphing offers appropriate and useful gradations in relation to what might count as autonomous and (perhaps eventually) sentient robotic systems. In fact, these perspectival gradations offer clear theoretical distinctions that can be repurposed or reconciled with robotic approaches to AGI in ways that might capture the interesting but currently-overlooked level of description that PSM offers. Combining PSM-levels of analysis in AI work with sociomorphing helps, also, to capture the asymmetry attached to social interactions including robots, an issue hinted at but never fully explored or explained in earlier social cognition work (De Jaegher et al., 2010)[2]. Additionally, while our goal here is to present a synthesis of these two approaches and show how they specifically can work in service of the AGI project that neither is individually aimed at, we also recognize that targeted experimentation within this framework, like all frameworks, is the next step to making sense of this problem and measuring the degree to which this new conceptual toolbox produces fruitful results. Therefore, our task here is to iteratively explore these concepts, and evaluate the terminological innovation as a sort of act of conceptual engineering in service of making sense of sense-making in concert with non-humans.

We recognize that the structure of this paper is a bit non-standard. We believe that this is necessary in order to incorporate the jargon of multiple niche academic areas in a way that will be understandable by naïve readers. To be clear, we understand our contributions to the literature to be as follows:

1) We borrow the concept of PSM from enactive social cognition and show how it is a valuable framework for making sense of behavior between humans and robots, while recognizing that this was not the domain it was designed for.
2) We borrow the concept of Sociomorphing from human-robot interaction (HRI) and show how it can be applied beyond the social robotics it was designed for, enlarging the scope of robotics-based AGI projects to help avoid known problems with anthropomorphization.
3) We offer a new conceptual framework for robotics-based AGI projects with the goal of focusing on social interaction as a methodology and location to study primary cognition with either or both human/animal and/or robot agents.

As such, we implore the reader to stick with us as we attempt to make various vernacular jargons familiar enough to work with, and hopefully to further deploy in AGI research.

The general structure of the paper is as follows: in Section The Argument, we offer the argument as we understand it, explaining PSM and its potential role in AGI, along with empirical support for the claims, in Section Background: PSM. In Section Sociomorphing, we discuss what sociomorphing is and why the concept is needed, both in HRI as it was originally intended, and also in AGI as we are applying it. In Section Background: Concepts of Sociality, we discuss the ways non-humans have been conceived of in social interactions across a range of literatures from various disciplines. In Section New Language and Conceptual Engineering, we offer evidence that anthropomorphism has failed to properly fulfill the explanatory role for which it is intended. In Section Failures of Anthropomorphism, we begin to put the pieces together to combine PSM and sociomorphing in a more complex way. This leads us to Section Revising "Social Interaction With Robots" in which we solidify the nuances of sociomorphing as a process, and try to show how its use forces a revision of our understanding of social interaction with robots, with a focus on perspective-taking. In Section Phenomenology, we enrich this argument by showing how it has explanatory power to make sense of the phenomenological experiences described by various human interactions with non-humans. Then in Section Autonomy is Hard, we revisit earlier complications involving senses of autonomy that need revision to reconcile enactive social cognition with robotics. We close the paper with a discussion in Section Discussion.

# THE ARGUMENT

## Background: PSM

In the classic example, we are asked to imagine two people trying to pass in a hallway, and frustratingly mirroring one another instead of successfully fulfilling the intentions of each, which are to simply keep walking. As opposed to traditional cognitive approaches to social cognition, in which an interaction with another person tends to be just a special instance of in-the-head cognition (in the form of mindreading or simulation), within enactivism, social cognition is a richer and more fundamental kind of cognition, in which genuine new kinds of meaning-making are enabled. In articulating the enactive concept of social cognition, De Jaegher and Di Paolo (2007, 2008) name it "participatory sense-making" (PSM). With roots in biological autonomy, the most general claim is that some kinds of social interactions produce a kind of cognition/sense-making that

[2]We have in mind here Di Paolo et al.'s provocative assertion that, "Interactions are social as long as the autonomy of the agents is not dissolved… We do not restrict social interaction to the human species. As long as the terms of the definition can be verified, they can apply to cross-species interactions or interactions with robots that are autonomous in the sense intended" (2010, 443).

emerges in the dynamics of the interaction and cannot be reduced back to the intentions of the individual actors.

Cognizers have a consistent perspective of their world because of the precariousness attached to their self-organization and self-maintenance, which entails needs and constraints relevant at multiple levels of their identities. Think of it this way: I need food, and I also need a friend. "A social interaction is an autonomous self-sustaining network of processes in the space of relations between the participants, provided their autonomies are also sustained" (De Jaegher et al., 2018, 139). An interaction process, in particular one that's considered autonomous in the right way, has patterns of coordination and breakdown which parallel the needs and constraints relevant to an individual sense-maker. Meaning, also, emerges in the same way from those patterns of coordination and breakdown which necessarily incorporate and sometimes supersede the processes surrounding the two (or more) individual sense-makers involved. Not all interactions count as social, and those that do have self-maintaining tendencies. In other words, the people involved in the interaction coordinate part of the exchange, and the exchange itself feeds back and encourages them to further sustain or modify the interaction. Think again about the people trying to pass in the hallway.

Recently, it has been argued that AI and AGI work has failed to consider and include claims from enactivism broadly and PSM specifically (Zebrowski and McGraw, 2021). In rethinking autonomy and openness in light of the enactive framework, social cognition, especially in this form, is highlighted as a central faculty of AI that has been widely overlooked. PSM has opened up a new research program to pursue with respect to AGI (and social robotics work). Yet what constitutes a sense-maker, and what is needed to produce an interaction process which fits the criteria necessary to be considered an *autonomous* sense-making thing in itself remain unclear in some conditions, largely as a result of both the traditional problem of other minds and also the empirical facts about human social behavior. In other words, there remain problems in describing and determining what constitutes a genuine sense-making system, particularly in regard to robotics, without running into epistemological roadblocks complicated by human tendencies toward anthropomorphising, and conflicting ideas of autonomy itself between enactivism and robotics (e.g., Haselager, 2005). In some of this recent work it is pointed out that there is a longstanding lack of consensus around concepts of agency and autonomy in the robotics and AI communities, as well as within many different systems of philosophical analysis (Zebrowski and McGraw, 2021). Drawing on Haselager (2005) and Barandiaran and Moreno (2006) the authors say: "In the most uncomplicated sense, a system is understood within robotics to be autonomous when it is able to perform its work without oversight. A robot is generally understood to be autonomous in the relevant sense when it acts in a way that precludes any human from being in the loop, and (perhaps more controversially) when it does so in an unpredictable environment. But a human is often considered autonomous in a much more radical sense: human autonomy tends to point toward a kind of metaphysical claim of free will… The least controversial sense of human autonomy is one that is limited to the ability to set and pursue one's own goals (309). They go on to say, "Within the enactive and PSM literature, an autonomous system is simply a system under precarious circumstances whose processes work to generate and sustain many of those processes as a source of self-identity" (310). So while the question of autonomy is a messy one, especially considering different fields and domains having distinct language and concepts, these ideas are not irreconcilable. The most convincing reconciliation comes from Di Paolo (2003), when he asks, "how can we invest artfacts with a similar sense of meaning? Do we need to go all the way down to metabolism and self-production or is another solution possible?" (12). He argues, instead of relying on the prototypical case of enactive autonomy, that of processes of life, we can focus on "the mechanisms for acquiring a *way of life*, that is, with habits" (13). Imbuing artificial systems with something like Deweyan habits then becomes an example of a way forward in making sense of autonomy that isn't limited to living systems. We will return at length to discuss anthropomorphism as it relates to this question, since this is the part of the equation that can be dealt with empirically.

Importantly for our purposes, we want to emphasize that enactivism has a complicated history with AGI. It is often ignored for its starting point in biological autonomy (although taken up and overlapping in some ways with historical uses of cybernetics). However, thinking about social cognition at all, and enactive social cognition in particular, would be an invaluable addition to AGI projects. Because of the historical bias in which minds are thought of as private, internal structures, social cognition tends to be an afterthought, if it is thought of at all in AGI. What PSM offers us is not merely the internal flipped outward, but a recognition that the private internal mind was never the right starting point. Instead, as PSM shows us, our interactions with other agents, as well as with the world, are the right starting point for making sense of social, and even individual cognition.

Reconciling an enactive theory of social cognition with a range of AGI projects, however, is difficult in multiple ways. One notable way is that PSM carries with it some assumptions about the kinds of systems that can be meaning-makers within these interactions. While teasing this apart is a large part of our overall project in this paper, we want to highlight one particular issue that deserves notice, because it is under-explored in the literature. Because only certain kinds of (usually biological and social) creatures can be interactors in obvious ways within PSM, there is a kind of assumed symmetry within these interactions. If we are engaged in joint meaning-making, we must both be social and cognitive creatures in at least roughly the same way. This symmetry proves particularly tricky in working out human interactions with animals, as well as with non-biological systems like social robots or other artificial cognizers (should such a thing 1 day exist).

There is empirical support for many of the claims in PSM across multiple methodologies and levels of description, from modeling at the neural level all the way to full embodied action (Reed et al., 2006; Auvray et al., 2009; Candadai et al., 2019). For example, Reed et al. (2006) performed an experiment in which two people might be linked haptically in trying to solve a target acquisition task. In the experimental condition, the

two people were linked, and responded to one another's motor control systems in trying to acquire the target, although they were ignorant of whether they were in the experimental condition or not. Subjects either independently moved a handle to place a projected mark into the target (in the single condition) or were tethered together in solving the problem (in the dyad condition). The results are surprising, and a reminder of why group or collaborative work, while often frustrating, tends to produce better outcomes than working alone. The authors say, "…task completion times indicated that dyads performed significantly faster than individuals, even though dyad members exerted large task-irrelevant forces in opposition to one another, and despite many participants' perceptions that their partner was an impediment" (365). In spite of feeling frustrated by their partners' real and perceived hindrances, pairs were faster and more successful at the task. Individual intentionality cannot aim toward this end; it can only be reached through an interaction with another person [see also a discussion of this experiment in De Jaegher and Di Paolo (2008), 143–144]. Similarly, Auvray et al. (2009) designed an experiment using a one dimensional plane on a computer screen to show that human participants can consistently detect and distinguish the presence of another person from that of both a fixed object and one that is mobile, as well as one that's a lure, shadowing the other participant's actual position. Otherwise sensory-deprived participants were given haptic feedback when crossing one another's activity, and the same feedback when crossing a fixed or mobile object. Each participant was told to perform an action (click a mouse) when they believe they've encountered another living participant. The authors state, "When the trajectories of the avatars cross, *both* participants receive a stimulation… each participant then turns back, then they will meet again, and this pattern forms a relatively stable dynamic attractor" (11). Given the sensory motor dynamics of the interaction, and the patterns of activity which arise through active engagement with the other living participant, they tended to create "…joint strategies of mutual exploration" resulting in the participants finding each others' avatars more often than not. These studies suggest that when two sense-making systems interact with one another (in regard to completing a specific task), an interesting new level of description tends to emerge between those two agents, one that couldn't have emerged for just a single individual. Thus, we see PSM in action: at least two autonomous systems in interaction, producing a new autonomous system that is dynamic and responsive to the individual interactors, but not always in a predictable or desirable way. It is also suggested that dynamical systems tools can model and measure this system.

## Sociomorphing

Bracketing for a moment the phenomenological experience, we want to acknowledge that there are many interactions between humans and robots, as well as mundane interactions between humans and animals, that seem to involve genuine meaning-making. Due to limitations on the kinds of robotic systems we have at this point in history, the meaning-making is largely one-sided in those interactions, but the system in interaction may well be autonomous enough that it will soon, if it doesn't already, count as its own rudimentary kind of cognizing system in a PSM-style interaction. But this asymmetry in interaction requires serious attention, especially if we ever hope to make the leap away from biological autonomy as the only actual (or conceptual) possibility of meaning-making.

In recognizing a conceptual and terminological gap in HRI research, Seibt et al. (2020a) have argued that the persistent approach of analyzing robots through the lens of anthropomorphism is mistaken; in its place, they offer a new ontology that takes account of perspectives (both of participants and observers) with a focus on asymmetrical social interactions. They argue that anthropomorphism as a frame hinders our ability to make sense of and study human interactions with social robots (in particular) because we mistakenly believe we impose human capacities and characteristics on machines which do not have them. Instead, they argue that there are genuine social capacities in animal and robot systems, but that we do not yet have a framework for understanding those social capacities in any way other than imposing human capacities on them. We are always already aware of the non-human capacities in some of our social interactions, and we already make adjustments to our behaviors based on that awareness, which isn't fully captured by an anthropomorphic analysis ("e.g., one can undress in front of a dog without being ashamed") (Seibt et al., 2020a, 63). In other words, you don't treat the dog as a person with human social skills and capacities, but you automatically make different judgements about its role in your social world. This is likely also true of, say, a robot dog like Aibo, although not necessarily in the same ways that it's true of a biological dog. Rather than anthropomorphizing, this, then, is sociomorphing.

What sociomorphing adds to this picture is a way to sidestep some of the problems of enactive autonomy, by offering a new conceptual framework in which we can conceive of asymmetrically-distributed social capacities across different kinds of systems. One of us (Robin) has a dog who appears to engage in frustratingly social interactions, wherein she (the dog) will steal an object that has been placed deliberately but unsuccessfully out of her reach, and then bring it into view so that whoever is nearby will attempt to retrieve it from her. This usually looks like Robin chasing the dog around angrily and yelling futilely, the dog appearing to greatly enjoy this interaction, running more the angrier Robin gets. This would appear to count as a kind of PSM, insofar as there are two autonomous beings engaged together in a single act, both of whom appear to be frustrating the intentions of the other, but both of whom also continue to engage not despite, but because of that frustration. Robin wants to get the object without chasing the dog, and the dog wants to be chased and sees the object only of instrumental value in reaching that goal, and yet both are drawn into this game repeatedly. You can see how the language of intentionality, when projected onto the dog, is controversial and less than ideal. While the game feels and looks quite deliberate from the human side, it is extremely difficult to project intentions onto a dog like this without running afoul of so much work in cognitive ethology. What sociomorphing offers is a new category of explanation, a new set of tools and language embedded in a whole new framework, by which we are already engaged in treating the other

with respect to its actual capacities and not imagined human capacities we know it doesn't possess. We anticipate different kinds of responses from a dog, or a robot, than we would from another person, and we react in the situation to those *actual* social capacities, not as if the other is capable of the narrow kinds of human interaction that anthropomorphism seems to demand.

## Background: Concepts of Sociality

While PSM offers a valuable lens through which to examine social interactions, particularly those involving humanoid robots or potential future AI systems, there have long been questions about the role of non-humans in such interactions. In examining the concepts of social interaction and social cognition, De Jaegher et al. (2010) introduce the possibility of robots being genuine social interactors under the right conditions. To reiterate, they say, "We do not restrict social interaction to the human species. As long as the terms of the definition can be verified, they can apply to cross-species interactions or interactions with robots that are autonomous in the sense intended" (443). We are left wondering how autonomy "in the sense intended" can happen in non-living systems, given that the sense intended generally centers processes of life and autopoiesis. The authors gesture toward future research questions, many of which overlap with our own here, including explorations of the characteristics of asymmetric social interactions and observational social understanding (such as watching a movie) (446).

There is no interdisciplinary consensus on the role of non-humans in social interaction, although reviews of the literature have been undertaken in several disciplines with stakes in the answer. Cerulo (2009) offers a broad review of theories mainly in sociology, but with reach into philosophy and psychology, too, which considers ways in which various forms of non-humans might fit into social interaction. A theme that emerges in Cerulo's review is that non-humans have been considered potential social interactors across a wide variety of theories, playing different roles and having different constraints. Most importantly, what emerges from this literature review is that interaction processes as well as the entities that potentially contribute to them have consistently gone through revision in tandem with evolving theories and technologies. For example, both Nass and Turkle noted that we interact with certain technologies and objects in fundamentally social ways,and in some cases those objects are capable of actively evoking feelings of intersubjectivity in us as we interact with them (Cerulo 539–540). Owens and Cohen, on the other hand, suggested that we consider non-humans "doing mind" in social interaction, understanding the non-human interactant as an other, treating it as independent of oneself and acting as if it has the capacities it seems to have (536). Importantly, this is not just projection and anthropomorphism, at least according to these theorists, indicating that this debate predates questions about AGI and social cognition.

In large part, Cerulo suggests that nonhumans in general "…deserve a more central place in our analytic frame" (543). She posits that the role and function of the mind in social interaction has been lacking, and that shifting focus to a more inclusive frame will help to fold in "…entities capable of different states of mind" (543). She claims a dog, for example, has been shown to have some awareness not only of "interactive routines" but also to establish "…cognitive, affective, and behavioral presence in interaction…" (Cerulo 544). This seems undeniably true in the case of Robin and her dog, but what of someone and their Aibo? Where the traditional (mostly sociological) theories Cerulo puts forward in her literature review have attempted to incorporate (in some form) nonhumans in interaction, they've not been definitive, in particular because social robotics and humanoid robots pose new kinds of problems, many of which relate to and are exasperated by failures of anthropomorphism.

Of course, philosophers of technology and others in human-robot interaction have long tackled this same question of sociality, asking what role the robot does or can play in our social interactions. Mark Coeckelbergh, in a paper laying out a new way to approach human-robot relations, discusses technology as an "other in itself." This work shifts the conversation away from what an entity *is* to what it *appears to be* (Coeckelbergh, 2011). The author uses Idhe's concept of an alterity relation to strip away the idea that an entity must have some prerequisite form of intentionality or consciousness to be considered properly social in human-technology relations. He sidesteps the ontological unknowns by focusing on "the appearance of the robot as experienced by the human" (199). Although Coeckelbergh's framing is phenomenological rather than sociological, like Cerulo's, the idea here links to an overarching conversation touching on anthropomorphism's cross-disciplinary relevance, and the mistakes therein. There is a large body of scholarly literature in both social robotics and social cognition that attempts to tackle these questions, but because there remains no cohesive HRI field that captures each of the relevant disciplines, the problem remains (Hakli and Seibt, 2017). A centralized, cross-disciplinary framework is necessary to at once sync fractured theories together and also to move toward a more cohesive tool in understanding the complexity surrounding social interactions, in particular asymmetrical ones. For our argument, this is a prerequisite to us being able to use this new combined framework and conceptual scheme described here.

The scholarly literature on non-humans in sociality shows the limits of anthropomorphism in making sense of social interaction broadly. Even the phenomenologically based relational views put forward by Coeckelbergh and others focus heavily on a human's experience of a social robot through an analysis shaped in terms of anthropomorphism. To this end, Seibt et al. introduce the concept of sociomorphing, which is the perception of real social capacities and characteristics, but not framed in terms of human sociality, which we have taken up here.

# NEW LANGUAGE AND CONCEPTUAL ENGINEERING

## Failures of Anthropomorphism

As mentioned, it has been argued that the field of AGI could benefit from the conceptual approach offered in PSM (Zebrowski and McGraw, 2021). Briefly, this is true because the enactive approach to social cognition offers new potential levels of cognitive analysis through the form of emerging

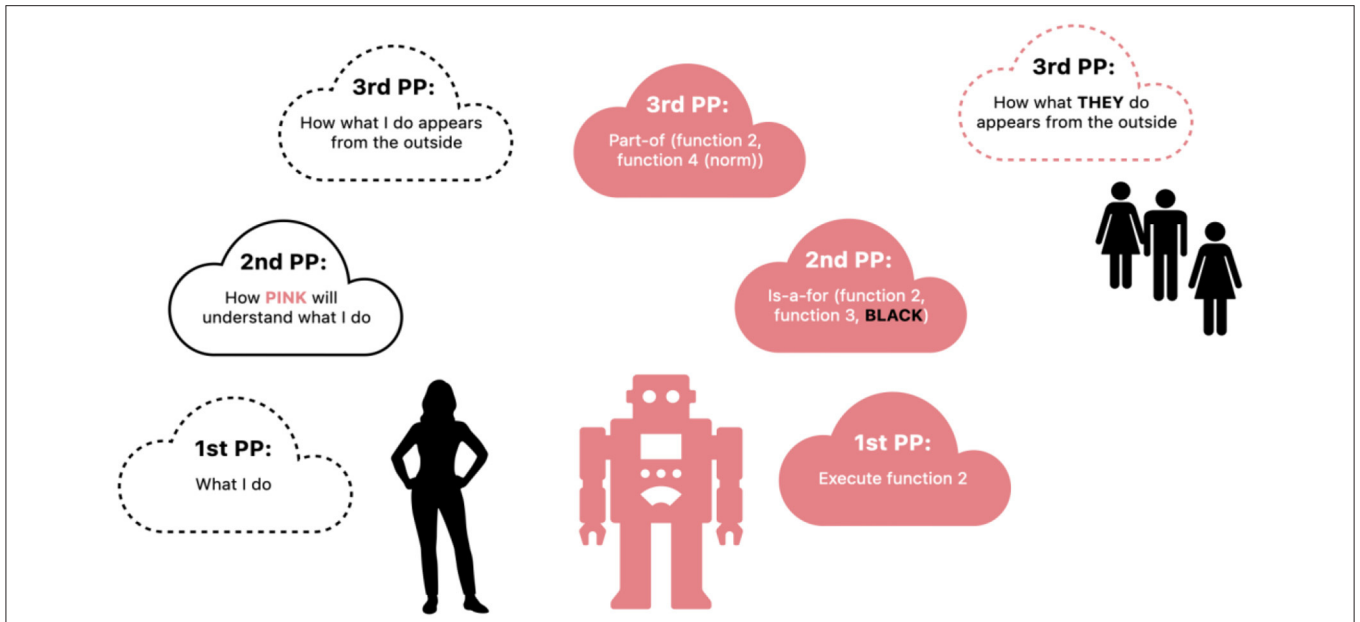



**FIGURE 1 |** The modified second-person perspective in asymmetrical social interactions, adapted by Eli McGraw and Jacqueline McGraw from Seibt, Vestergaard and Damholdt 2020, with permission.

dynamic systems in social interaction. This joins historical calls to center 4e cognition and embodiment, including humanoid embodiment, in AGI (Holland, 2007; Chella et al., 2019). Zebrowski and McGraw (2021), which we build on here, centers on reconciling enactivism's biologically-based notion of autonomy with muddier notions of autonomy used across robotics and AGI, as mentioned in Section Background: PSM. Using the conceptual and empirical tools offered by PSM, progress might be made toward understanding the possibilities enabled by certain kinds of artificial systems in certain kinds of social interactions with certain kinds of living systems. However, this picture leaves open the possibility of perceptually-indistinguishable but ontologically-distinct pictures, like the traditional philosophical zombie problem.

If PSM is going to be a valuable framework for understanding social robotics and (more importantly for our purposes) future AGI work, then we need to understand the role of the non-human system in that interaction better than any framework currently does. In spite of an exploding literature in human-robot interaction (HRI) and social robotics as they relate to anthropomorphising, Seibt et al. newly suggest that anthropomorphism isn't actually all we're doing when we as humans engage with (social) robots. In other words, part of the roadblock in making sense of social interactions with robots broadly, and within enactive frameworks specifically, has been a lack of terminology to make proper sense of the ontological status of each participant. Sociomorphing, then, is a "terminological innovation" that provides us with conceptual and empirical approaches not previously accessible within this framework.

The traditional accounts of social interaction imply or explicitly demand that all interactants have some number of certain kinds of (human) social capacities, including consciousness, intentionality, self-awareness, empathy, emotions, beliefs, reasoning, capacity for joint-action, etc. (Duffy, 2003; Cerulo, 2009; Hakli, 2014; Parviainen et al., 2019; Damholdt et al., 2020; Seibt et al., 2020a). With regard to human-robot interaction (often social robotics), the literature on anthropomorphism has always been contentious (Duffy, 2003; Waytz et al., 2010; Darling, 2017; Epley, 2018; Zebrowski, 2020). Many researchers point out that our projection of human capacities onto non-human systems results in a *metaphorical* use of anthropomorphism already. Parvianen et al., for example, say "currently, the robot functions are described metaphorically in the human-robot interaction literature, which refers to human consciousness capabilities. Robots are said to "sense," think," and "act" (Parviainen et al., 2019, 323). Duffy (2003) points out that anthropomorphism "includes metaphorically ascribing human-like qualities to a system based on one's interpretation of its actions… [it] is a metaphor rather than an explanation of a system's behavior" (181). Within social robotics in particular, where the systems are designed to look social and appeal to evolutionary responses people may have that judge certain traits as social, we can easily see the mismatch between the concept of anthropomorphism and its application.

One way that anthropomorphism seems to fail as a proper frame in human-robot interactions is that there is a wide gulf between the kinds of human capacities that produce certain behaviors in humans and those similar kinds of behaviors in robots (linguistic behavior here is the most obvious; when I say "I love you" it means something very different than when an Aibo or a Pepper robot says it. The same is true of many other imitative behaviors). There is an asymmetry in these interactions that causes a shift in perspective as to how I understand the robot's actions, and how I understand what the robot will

understand of my actions. This asymmetry isn't new, and it holds for many of our human-animal interactions, as well as interactions with humans who differ from us widely in age, class, or neurotypicality[3].

There has long been a call for new language both to conceptualize and study human interactions with robots without relying on anthropomorphism. Duffy (2003) paper includes a clear recommendation for this: "Anthropomorphism should not be seen as the 'solution' to all human-machine interaction problems but rather it needs to be researched more to provide the 'language' of interaction between man and machine" (181). Coeckelbergh (2021), also calls for an overhaul of our understanding of anthropomorphism as it relates to social robots. In writing about the use of social robotics in relation to social interaction, Hakli (2014) gestures to the shortcomings of sociality broadly, anthropomorphic tendencies more specifically, and the need to produce new language which takes into account the breadth of the social. He surveys various ways social robots have been excluded from the concept of sociality by definition, ruling them out by defining social interactions in terms of consciousness or intentionality. Instead, he argues that perhaps "if people conceive of their interaction with robots as social interaction, this should count as *prima facie* evidence that their interaction with robots is social interaction" (107). Conceptualizing sociality as fluid and malleable rather than having well-defined boundaries can completely alter our ways of understanding social cognition. These definitional pitfalls, to Hakli, shed light on the need to rethink the concepts that provide structure for our theoretical approaches in HRI and with asymmetric interactions more broadly. Hakli points to the fact that other theorists have undertaken the challenge to "…build conceptual frameworks with graded notions and several dimensions that enable us to locate more carefully the differences between different types of agents" (113). Seibt et al.'s project with regard to sociomorphing is an attempt to answer these calls for new language and concepts, and we hope to take this challenge up and apply it further into AGI work, too.

## Revising Social Interaction With Robots

On some level, we're asking how we can judge that a robot has developed a mind, but this framing misses all of the nuance needed when trying to make sense of something long misunderstood as a binary. While there's a simple way that this is nothing more than an illustration of the traditional problem of other minds, it seems that taking up (Seibt et al., 2020b) framework helps chisel away at more of the problem through clarifying and reconceptualizing some of the framework we superimpose in analyzing human-robot social interactions. In other words, what has the luxury of being a purely theoretical issue for philosophers is much more pressing for practitioners, and we hope to concretize some of this theory for researchers to take up in practice. In approaching these interactions as involving sociomorphing instead (or alongside) of anthropomorphizing, we shift our focus to the actual capacities of the artificial system instead of fictionalizing the interaction as if it were between just human interactors. When I interact with another person, I know that the social capacities for our interactions are more or less symmetrically distributed, and that guides the perspectives I take on such an interaction. But when we interact with some kinds of social robots or animals, the authors argue that part of what we do is generate a new model that tries to account for the asymmetry of the interaction. Sociomorphing involves all of the following (p. 58):

> "(S1) it is *direct perception*;
> (S2) it is a perception of *non-human* characteristics and capacities (which resemble certain characteristics and capacities familiar from human social interaction to different and possibly very low degrees) in non-human entities and circumstances;
> (S3) it both arises in and guides *interactive sense-making* in a situation of practical interaction (or the perception of an interaction);
> (S4) it typically occurs preconsciously but may also occur consciously; and
> (S5) it pertains to (relative to an external observer) *actual* features of non-human entities and characteristics."

We implicitly pick up on the shifting perspectives needed to move from symmetrical to asymmetrical social interactions, largely shifting the second-person perspective (see **Figure 1**). If we look at this figure, we can see the seven perspectival perceptions of an asymmetrical social interaction with a social robot as described by Seibt et al. In a default social interaction, for example between two peers, they sociomorph one another using that default second-person perspective with the assumption that "the capacities required for the present kind of social interactions are symmetrically distributed" (61). However, in this figure, we can see how the human, possibly implicitly, understands that the robot will take the human's actions differently than a peer would, and changes the interaction accordingly. My understanding of my own action is different when I consider (through the second person) how my actions are being perceived by a robot or a dog (for example). There would be a cascade of changes given the "non-default" ness of an asymmetrical situation. While the authors here point out seven specific perspectives involved in this sort of interaction, elsewhere they write that social interactions have "irreducible perspectival complexity" (Seibt, 2018, 140; Seibt et al., 2020a, 137). We suspect that the designers of the robot also need to be represented in this picture, as the human interacting with it will tend to default to what they think the system was designed to do, but detailing the richness of these perspectives is beyond the scope of this paper[4].

[3]An important note: In more than one place (2017, 2018), Seibt has laid out features of an entire ontology that tries to make sense of what she sometimes calls "simulated social interaction." This ontology, called OASIS (Ontology of Asymmetric Social Interactions) lays out a number of specific features of these interactions, classifying different kinds of human-robot coordinated and collaborative actions through different sorts of simulations. This ontology can be closely tied to different kinds of sociomorphing, but the specific levels of her ontology don't quite capture our use of her concept. This is likely because we are attempting to expand this idea beyond the framework of social robotics into AGI/humanoid robotics work, which hopes to escape the focus on simulation and eventually emerge as simply doing.

[4]Coeckelbergh, as one example, states "I argue that designers have a responsibility for designing the role and narrative related to their artifact, and indeed designing the performance, but that the performance and the narrative created are also the

## Phenomenology

Recall that the primary aim of sociomorphing is to recognize the perception of actual social capacities in a thing or person without assuming evenly-distributed symmetrical social skills. Along with this, the authors introduce a concept of types of experienced sociality (TES) understood as the feelings of co-presence with another agent or entity. They hypothesize that "sociomorphing can take many forms each of which is manifested in, or otherwise associated with, a *type of experienced sociality*" (Seibt et al., 2020a, 52) (emphasis in original). New TES's occur when one agent "operates with a 2nd person perspective that deviates from the symmetry assumption" (61). A portion of any given TES touches on a "feeling of co-presence or 'being-with'" a particular social interactor (59). Imagine, for example, the phenomenological distinction between what it feels like to be in a shared space with a dog, or cat, vs. with a Pepper robot or another human being. They acknowledge that in some cases, these capacities may be perceptually indistinguishable from corresponding human capacities (making this an ontological claim, but one that can be tested or deployed in empirical settings to make finer distinctions within the research). The feeling of co-presence is relevant in each case, but the type of experienced sociality changes given differing expectations of the situation, including things like anticipated responsive capabilities and environmental circumstances. Some forms of sociomorphing and the TES's associated with them can be mapped onto others intuitively. The TES of being-with PARO or an Aibo for example, might resemble that of being-with a cat or dog, or it may not at all, depending on the context of the interaction. These complexities which arise within two conceptually similar TES's point to the idea that a new descriptive framework is needed to better anchor various forms of sociomorphing, and the shifting asymmetries across different (actual and potential) sense-making systems.

Including this phenomenological feature, the TES, in the ontological picture is long overdue. The experience of being-with certain kinds of animal-like or human-like robots has long been reported as similar to being-with a being. For example, Turkle (1995) reported her first experience with MIT's Cog as having been surprising, since she knew what the robot was and what it was (not) capable of. But still, she says, "Cog 'noticed' me soon after I entered its room. Its head turned to follow me and I was embarrassed to note that this made me happy. I found myself competing with another visitor for its attention… Despite myself and despite my continuing skepticism about this research project, I had behaved as though in the presence of another being" (266). Similarly, Darling (2021) reports a surprising experience when visiting Boston Dynamics. She, like Turkle, had plenty of experience with the generally non-functioning robots in the labs at MIT, and when she saw one non-functioning robot slumped over at Boston Dynamics, she remarked that people often think the robots are more functional than they ever actually are. But she goes on, "My jaw dropped. Behind the door was a gymnasium-sized hall outfitted with an elaborate obstacle course. Dozens of dog-sized robots were roaming the premises, walking up and down stairs, pacing back and forth in pens, or ambling around the area completely by themselves" (Darling, 2021, 102). In spite of their rich familiarity with robots, each of these researchers reports a surprising TES, a phenomenal experience of being-with another kind of living being. This phenomenon is not new and has been widely explored in relation to phenomenology (Zebrowski, 2010). Darling's language clearly draws on the experience of being-with dogs, but it remains unclear how much she sociomorphs the Boston Dynamics robots as dogs and how much she instead implicitly takes up a similarly non-default robotics-based perspective in this interaction. For this reason, assumptions and intuitions about the TES and associated variety of sociomorphing cannot be determined in advance of targeted research within this framework. Given this, both empirical and theoretical researchers in HRI and AI ought to take up the sociomorphing framework to better understand how people without as much experience with robots as, for example, Turkle and Darling do, will experience them as animals or people, and what role these systems can play in social sense-making.

Our TES associated with being-with a dog or baby is likely repurposed with a social robot until the robot speaks (like Sony's Aibo); imagine encountering the fictional Lying Cat from the comic *Saga* (Vaughn and Staples, 2012). (Lying Cat is a large bluish cat in the comic, who speaks one word, "lying" when someone is lying. His metaphysical capacity to always know this is never explained, and reflects no counterpart in the actual world we live in). He, too, is a kind of categorical novelty in the same way as the social robot, and sociomorphing captures the experience in a way anthropomorphizing does not. The capacity of a large cat to be a sense-maker in an interaction process depends in part upon the way it fits my preconceived idea of what social action looks like through my own lived experience. Importantly, though, I am already sociomorphing and considering a new TES that goes beyond the scope of any previous interactions in such a way that introduces the perception of the manifestation of a new capacity (because cats have never done this). The joint process that arises between myself and Lying Cat then, can be considered in terms of a new *something* not previously available explicitly within the enactive social model of PSM. Therefore, the sociomorphing framework enriches our application of PSM to AGI specifically, and not just HRI as proposed by the authors. We again suggest that this enlargement of this framework has explanatory power beyond interaction between humans and simulation-based systems.

# AUTONOMY IS HARD

We began this paper claiming that PSM is an overlooked theoretical approach to social cognition that researchers in AGI would be wise to take up. This is the case because it describes the social interaction in terms of autonomous systems, which, in proper kinds of interaction, generate new autonomous systems capable of being studied on their own. It denies the "rear-window" approach to social cognition, refusing to allow in-the-head intentionality to be the appropriate level of description as it

responsibility of the user since they also co-create it" (Coeckelbergh, 2018, 73). This kind of feedback loop results in complex perspective-taking that we seem able to perform and create easily, but the theorizing behind such complexity is not nearly as easy.

remains in both mindreading and simulation theories. It explains empirical data that shows how multiple people in interaction perform in ways that are more than the sum of their parts, as well as offering a richer, enactive view of cognition that doesn't merely apply individual cognition to two or more people.

PSM, though, is a system requiring multiple levels of autonomy: the individual interactors as well as the emergent dynamics under the right conditions. At the core of these autonomies and interactions is what Di Paolo et al. call a "primordial tension." This is not a tension between two agents in interactions, but between the agents and the situation in which PSM emerges. This tension persists "even if others are not present *as others* or if there is no discord between intentions at all" (140). In other words, even if you aren't aware that some of your interactions are with another cognitive system, PSM may still occur. While we've avoided saying too much about autonomy itself here, we must point out that autonomy in the enactive sense is based in life processes, or at least a kind of self-sustaining system which has not yet been achieved in artificial systems. Autonomous robots, as the term is used, share very little with autonomous systems within the enactive framework, and we want to be careful not to equivocate to solve the problem. However, as mentioned in Section Background: PSM, Di Paolo (2003) has posed a potential way to resolve the conflict by refocusing enactive autonomy away from life processes and towards a "*way of life*" (13), a difference that can make a difference in actual robotics-based AGI research. Seibt et al.'s framework of sociomorphing and types of experienced sociality (TES) help us reconceive the interaction process in a way that opens up possibilities of gradations in how we understand autonomy. Instead, the interaction is understood as a new kind of thing, that requires new perspective taking. This reframing allows for new ways of thinking about interacting with agents that may or may not (yet?) be autonomous in the sense intended.

For example, in the Reed et al. (2006) study discussed earlier, two people are interacting to solve a task, in spite of being frustrated by the interaction process (indeed, people were unaware of which condition they were in, the single condition or the dyadic condition, and hence were unaware of the other *as other in interaction)*. The interaction, in this case, enables the cognition needed to solve the problem quicker and more accurately than either could solve it alone. The interaction process, in this case, seems to, "deliver the necessary cognitive performance" (De Jaegher et al., 2010). In other words, social cognition is not just enabled here, but constituted in and by the interaction process itself, as PSM predicts. As summed up by De Jaegher et al. (2018) "in cases of synergy between individual and interactive normativity, acts acquire a magic power. They achieve more than I intend to" (143). We are arguing that this magic power may actually emerge among other systems of sociality, including possible or actual emergent dynamics between human and artificial systems.

Take Sony's Aibo robot dog, for example. In 2015, the New York Times documented the phasing-out of Aibo, and the impact it was having on multiple Japanese families who were holding funerals as the Aibos became unrepairable. Aibo was a categorical novelty insofar as it was a robot, but it looked (sort of) like a dog, and it was trainable, so each person's interactions with their Aibo would change the system to optimize future interactions with that particular owner or family. Many of the owners were empty-nesters who took the robot on as a member of the family. As mentioned earlier, the TES of being-with Aibo may or may not be like being with a dog, or it may be a bit like being with a baby. One of the owners says, "When I first got Aibo it was like having a new baby. It wasn't just a robot, because we had to raise it" (The New York Times, 2015). In terms of anthropomorphism, we might analyze this phenomenon as if the robot dog was replacing a biological dog, or a child no longer living at home. But neither of these explanations captures the actual role of the robot dog, or captures the way the owners interacted with their Aibos (shown on an untranscribed video). To the owners, their feelings of co-presence associated with their prior experiences of dogs or babies only brought them so far in attempting to capture their understanding of Aibo as an other. Here, the owners' second person perspective was reshaped given the asymmetry of the process at hand (see **Figure 1** again). They implicitly and explicitly realized the fact that Aibo had capacities different from (but including some) of those of another peer, a dog, or a baby. Aibo had capacities of its own, and the owners treated it as a social interactor in its own right. The Aibo-owners' expectations and interactions with their robot dogs are a bit like the inverse of Darling and Turkle's interactions described above, at least at first glance. But in reality, the expectations of anthropomorphism aren't enough to capture the actual interactions between human and robot-as-social actor that we see in each case. Aibo's owners are already engaged in sociomorphing, but without the conceptual framework and language to describe it. Aibo clearly lacks autonomy in the enactive sense, but as new technologies emerge that might properly engage in social interactions with humans, we need a finer distinction in how we understand and experience sociality. And while there are no robots autonomous in the enactive sense (yet), perhaps gradations of autonomy are already emerging (the dog-like robots at Boston Dynamics described by Darling above may be an interesting candidate) and this framework guides us toward more productive design of these systems.

## DISCUSSION

By combining the approaches of both PSM and sociomorphing, we have a new way to empirically study and theoretically examine artificial systems as they exist in interaction with us without constantly encountering epistemological roadblocks of systems built to *appear* social. In fact, Seibt and her collaborators have offered an in-depth ontology alongside the sociomorphing framework that attempts to provide descriptive tools related to different levels of simulation and asymmetry, carving up the problem space into at least five different kinds of simulation (Seibt, 2017; Seibt et al., 2020a). They have also created a new instrument meant to take the theoretical structures into the experimental sciences (Damholdt et al., 2020). Both traditional HRI research, as well as AGI work, are in need of this new terminology and framework, especially in light of the level

of description PSM provides in studying social cognition. For example, in terms of Lying Cat, sociomorphing introduces a finer-grained approach to analyzing the actual systems at hand. In other words, we don't live in a universe with talking cats who have access to metaphysical truths, so if I interact with an actual cat who seems to be uttering "lying," the framework of sociomorphing that I use appropriately leaves out linguistic capacities from the analysis. In the universe of *Saga*, that would be different. In the case of an actual dog vs. Aibo, anthropomorphism relies on language that may incorrectly equate the two, which doesn't allow for the literal ascription of social capacities onto either. Sociomorphing allows for us to take each of those systems as genuinely social, without only projecting human capacities onto them. In fact, if we take up sociomorphing broadly, other asymmetrical social interactions we regularly engage in (such as when we say hi to the neighbor's dog or while watching birds out the window) would be reframed as well, and we would have language and conceptual systems that accurately capture the real cognitive and emotional states of those systems. If each form of sociomorphing manifests in a TES, and TES is fundamentally tied to phenomenological experiences of co-presence as well as perspectival shifts, then AGI needs a reckoning with the overall framework used to analyze cognition broadly, but more specifically social cognition, most productively understood as PSM.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## REFERENCES


Auvray, M., Lenay, C., and Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas Psychol.* 27, 32–47. doi: 10.1016/j.newideapsych.2007.12.002

Barandiaran, X., and Moreno, A. (2006). On what makes certain dynamical systems cognitive: a minimally cognitive organization program. *Adapt. Behav.* 14, 171–185. doi: 10.1177/105971230601400208

Candadai, M., Setzler, M., Izquierdo, E., and Froese, T. (2019). Embodied dyadic interaction increases complexity of neural dynamics: a minimal agent-based simulation model. *Front. Psychol.* 10:540. doi: 10.3389/fpsyg.2019.00540

Cerulo, K. (2009). Non-humans in social interaction. *Annu. Rev. Sociol.* 35, 531–552 doi: 10.1146/annurev-soc-070308-120008

Chella, A., Cangelosi, A., Metta, G., and Bringsjord, S. (2019). Editorial: Consciousness in humanoid robots. *Front. Robot. AI.* 6:17. doi: 10.3389/frobt.2019.00017

Coeckelbergh, M. (2011). Humans, animals, and robots: a phenomenological approach to human-robot relations. *Int. J. Soc. Robot.* 3, 197–204. doi: 10.1007/s12369-010-0075-6

Coeckelbergh, M. (2018). How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics Inf. Technol.* 20, 71–85. doi: 10.1007/s10676-017-9441-5

Coeckelbergh, M. (2021). Three responses to anthropomorphism in social robotics: towards a critical, relational, and hermeneutic approach. *Int. J. Soc. Robot*. doi: 10.1007/s12369-021-00770-0

Damholdt, M., Vestergaard, C., and Seibt, J. (2020). "Testing for 'Anthropomorphization': a case for mixed methods in human-robot interaction," in *Human-Robot Interaction*. Springer series on bio- and neurosystems, Vol. 12, eds C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, et al. (Cham: Springer Nature), 203–227. doi: 10.1007/978-3-030-42307-0_8

Darling, K. (2017). "Who's Johnny' anthropomorphic framing in human-robot interaction, integration, and policy," in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, eds P. Lin, R. Jenkins, and K. Abney (Oxford: Oxford University Press), 173–188. doi: 10.1093/oso/9780190652951.003.0012

Darling, K. (2021). *The New Breed: What Our History With Animals Reveals About Our Future with Robots*. New York, NY: Henry Holt.

De Jaegher, H., Cuffari, E., and Di Paolo, E. (2018). *Linguistic Bodies: The Continuity Between Life and Language*. Cambridge, MA: MIT Press.

De Jaegher, H., and Di Paolo, E. A. (2007). Participatory sense-making, *Phenomenol. Cogn. Sci*. 6, 485–507. doi: 10.1007/s11097-007-9076-9

De Jaegher, H., and Di Paolo, E. A. (2008). "Making sense in participation: an enactive approach to social cognition," in *Enacting Intersubjectivity*, eds Morganti, F., Carassa, A., and Riva, G. (Amsterdam: IOS Press), 33–47.

De Jaegher, H., Di Paolo, E. A., and Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends Cogn. Sci.* 14, 441–447. doi: 10.1016/j.tics.2010.06.009

Di Paolo, E. (2003). "Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop," in *Dynamic Systems Approach for Embodiment and Sociality: From Ecological Psychology to Robotics*. eds Murase, K., and Asakura, T. (Advanced Knowledge International), 19–42.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robot. Auton. Syst.* 42, 177–190. doi: 10.1016/S0921-8890(02)00374-3

Epley, N. (2018). A mind like mine: the exceptionally ordinary underpinnings of anthropomorphism. *J. Assoc. Consum. Res.* 3, 591–598. doi: 10.1086/699516

Fuchs, T. (2021). "Human and artificial intelligence: a clarification," in *Defense of The Human Being*. eds Newell, Simon, and Dennis, W. (Oxford: Oxford University Press), 13–48. doi: 10.1093/oso/9780192898197.003.0002

Hakli, R. (2014). "Social robots and social interaction," in *Sociable Robots and the Future of Social Relations*, eds J. Seibt, R. Hakli, and M. J. Nørskov (Amsterdam: IOS Press), 105–114.

Hakli, R., and Seibt, J. (eds.) (2017). *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interactions*. Cham: Springer. doi: 10.1007/978-3-319-53133-5

Haselager, W. (2005). Robotics, philosophy and the problems of autonomy. *Pragmat. Cogn.* 13, 515–532. doi: 10.1075/pc.13.3.07has

Holland, O. (2007). A strongly embodied approach to machine consciousness. *J. Conscious. Stud.* 14, 97–110.

Newell, A., and Simon, H. A. (1961). "The simulation of human thought," in *Current Trends in Psychological Theory* (Pittsburgh, PA: University of Pittsburgh Press), 152–179.

Parviainen, J., Van Aerschot, L., Särkikoski, T., Pekkarinen, S., Melkas, H., and Hennala, L. (2019). Motions with emotions? A phenomenological approach to understanding the simulated aliveness of a robot body. *Techné: Res. Philos. Technol.* 23, 318–341. doi: 10.5840/techne201911 26106

Reed, K. B., Peshkin, M., Hartmann, M. J., Grabowecky, M., Patton, J., and Vishton, P. M. (2006). Haptically linked dyads: Are two motor control systems better than one? *Psychol. Sci.* 17, 365–366. doi: 10.1111/j.1467-9280.2006.01712.x

Seibt, J. (2017). "Towards an ontology of simulated social interaction: varieties of the 'As If' for robots and humans," in *Sociality and Normativity for Robots*, eds Halki, R., and Seibt, J. (Cham: Springer), 11–39. doi: 10.1007/978-3-319-53133-5_2

Seibt, J. (2018). "Classifying forms and modes of co-working in the Ontology of Asymmetric Social Interactions (OASIS)," in *Envisioning Robots in Society-Power, Politics, and Public Space*, eds M. Coeckelbergh, J. Loh, M. Funk, J. Seibt, and M. Nørskov (Amsterdam: IOS Press), 133–146

Seibt, J., Damholdt, M., and Vestergaard, C. (2020a). Integrative social robotics, value-driven design, and transdisciplinarity. *Interact. Stud.* 21, 111–144 doi: 10.1075/is.18061.sei

Seibt, J., Vestergaard, C., and Damholdt, M. (2020b). "Sociomorphing, not anthropomorphizing: towards a typology of experienced sociality," in *Culturally Sustainable Social Robotics – Proceedings of Robophilosophy 2020*, eds M. Nørskov, J. Seibt, and O. S. Quick (Amsterdam: IOS Press), 51–67. doi: 10.3233/FAIA200900

The New York Times (2015). *Robotica*, ep. 6. Available online at: https://www.nytimes.com/2015/06/18/technology/robotica-sony-aibo-robotic-dog-mortality.html (accessed October 18, 2021).

Turkle, S. (1995). *Life on the Screen: Identity in the Age of the Internet*. New York, NY: Simon and Schuster.

Vaughn, B. K., and Staples, F. (2012-present). Saga. Image Comics.

Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* 5, 219–232. doi: 10.1177/1745691610369336

Zebrowski, R. (2010). In dialogue with the world: Merleau-Ponty, Rodney Brooks, and embodied artificial intelligence, *J. Consciou. Stud.* 17, 156–172. Available online at: https://www.ingentaconnect.com/contentone/imp/jcs/2010/00000017/f0020007/art00009

Zebrowski, R. (2020). Fear of a bot planet: anthropomorphism, humanoid embodiment, and machine consciousness. *J. Artificial Intelligence Conscious.* 7, 119–132. doi: 10.1142/S2705078520500071

Zebrowski, R., and McGraw, E. (2021). Autonomy and openness in human and machine systems: participatory sense-making and artificial minds. *J. Artificial Intelligence Conscious.* 8, 303–323. doi: 10.1142/S2705078521500181

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Zebrowski and McGraw. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

HYPOTHESIS AND THEORY
published: 15 June 2022
doi: 10.3389/fnbot.2022.843108

# Grounding Context in Embodied Cognitive Robotics

*Diana Valenzo*[1], *Alejandra Ciria*[2], *Guido Schillaci*[3] *and Bruno Lara*[1]*

[1] *Laboratorio de Robótica Cognitiva, Centro de Investigación en Ciencias, Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico,* [2] *Facultad de Psicología, Universidad Nacional Autónoma de México, Mexico City, Mexico,* [3] *Independent Researcher, Florence, Italy*

Biological agents are context-dependent systems that exhibit behavioral flexibility. The internal and external information agents process, their actions, and emotions are all grounded in the context within which they are situated. However, in the field of cognitive robotics, the concept of context is far from being clear with most studies making little to no reference to it. The aim of this paper is to provide an interpretation of the notion of context and its core elements based on different studies in natural agents, and how these core contextual elements have been modeled in cognitive robotics, to introduce a new hypothesis about the interactions between these contextual elements. Here, global context is categorized as agent-related, environmental, and task-related context. The interaction of their core elements, allows agents to first select self-relevant tasks depending on their current needs, or for learning and mastering their environment through exploration. Second, to perform a task and continuously monitor its performance. Third, to abandon a task in case its execution is not going as expected. Here, the monitoring of prediction error, the difference between sensorimotor predictions and incoming sensory information, is at the core of behavioral flexibility during situated action cycles. Additionally, monitoring prediction error dynamics and its comparison with the expected reduction rate should indicate the agent its overall performance on executing the task. Sensitivity to performance evokes emotions that function as the driving element for autonomous behavior which, at the same time, depends on the processing of the interacting core elements. Taking all these into account, an interactionist model of contexts and their core elements is proposed. The model is embodied, affective, and situated, by means of the processing of the agent-related and environmental core contextual elements. Additionally, it is grounded in the processing of the task-related context and the associated situated action cycles during task execution. Finally, the model proposed here aims to guide how artificial agents should process the core contextual elements of the agent-related and environmental context to give rise to the task-related context, allowing agents to autonomously select a task, its planning, execution, and monitoring for behavioral flexibility.

**Keywords: context, behavioral flexibility, task selection, prediction error, cognitive robotics**

OPEN ACCESS

***Edited by:***
*Adam Safron,*
*Johns Hopkins Medicine,*
*United States*

***Reviewed by:***
*Jun Tani,*
*Okinawa Institute of Science and Technology Graduate University, Japan*
*Matthew Sims,*
*Ruhr University Bochum, Germany*

***Correspondence:**
*Bruno Lara*
*bruno.lara@uaem.mx*

***Received:*** *24 December 2021*
***Accepted:*** *10 May 2022*
***Published:*** *15 June 2022*

***Citation:***
*Valenzo D, Ciria A, Schillaci G and Lara B (2022) Grounding Context in Embodied Cognitive Robotics. Front. Neurorobot. 16:843108. doi: 10.3389/fnbot.2022.843108*

## 1. INTRODUCTION

Cognitive robotics (CR) aims to understand cognition by recreating it in artificial agents (Asada et al., 2001; Krichmar, 2012; Cangelosi and Schlesinger, 2015; Lara et al., 2018). In doing so, the interaction with the environment is assumed to be crucial for the emergence of cognitive abilities (Pezzulo et al., 2011, 2013; Cangelosi et al., 2015). Artificial agents are considered as useful tools to explore embodied, embedded, and grounded models of cognition (Pfeifer and Scheier, 2001; Lungarella et al., 2003; Pfeifer, 2004). Here, grounded cognition is understood as a general approach that incorporates embodied, embedded, enactive, and extended cognition into a broader perspective: "cognition, affect, and behavior emerge from the body being embedded in environments that extend cognition, as agents enact situated action reflecting their current cognitive and affective states" (Barsalou, 2020b, p.2).

Artificial agents are able to explore and manipulate objects in their environments (Min et al., 2016; Adnan Mohsin Abdulazeez, 2021). However, these tasks are usually learned under controlled conditions, which restricts their ability to efficiently adapt to the demands of dynamic environments (Min et al., 2016). One of the great challenges in Cognitive Robotics (CR) is to design autonomous artificial agents that generate appropriate behaviors according to the environment in which they are situated (Mohan et al., 2013; Asada, 2020). A promising approach is the attempt to understand the underlying mechanisms of behavioral flexibility that biological agents naturally exhibit. Behavioral flexibility refers to the ability to switch from one behavior to another so as to efficiently adapt to dynamic environments (Ragozzino, 2007; Lea et al., 2020). In this regard, context processing plays an essential role in behavioral flexibility.

The processing of the current context is fundamental for biological agents to select the appropriate task at a given moment. It is widely accepted that context acts as a set of constraints that influence behavior (Bazire and Brézillon, 2005). Actually, it makes no sense to talk about appropriate behaviors without the notion of context (Turner, 1998). Furthermore, contextual information is also essential for planning the sensorimotor sequences to execute a selected task (Rosenbaum et al., 2014). It has been suggested that the brain is a context-dependent system since all inputs it processes concern the context in which they occur (Nikolić, 2010). Following this line, processing context would allow artificial agents to autonomously and appropriately prioritize goals, select appropriate tasks, plan and execute them, and even change tasks according to the current situation, ultimately showing greater behavioral flexibility.

This paper aims to analyze the role of context in behavioral flexibility and how this concept has been used in CR. Although context is a widely used concept, not only in CR but also within cognitive sciences in general, it remains an ill-defined concept (for an attempt to analyze different definitions of the concept of context see Bazire and Brézillon, 2005). Inspired by the pioneering work of Turner (1998) in context-mediated behavior for artificial agents, here, context is defined as any identifiable configuration of environmental, task-related, and agent-related elements that are perceived and experienced as relevant in a specific moment and in a particular situation. To respond to changing conditions, biological agents must monitor internal demands and environmental factors, those that are of self-relevance and full of affect, to guide and initiate behavior (Barsalou, 2020b). Together, all those internal and external elements of a situation that have predictive power and impacts behavior constitute the global context (Turner, 1998, p.308). In order to unravel the diffuse notion of context and considering the key constituents of the definition proposed by Turner (1998), three components of the global context are considered in our analysis: agent-related, environmental, and task-related context (**Figure 1**). Pfeifer and Bongard (2006) considered the same components within their set of design principles for artificial agents, stating that an intelligent agent should have a defined ecological niche, a defined task, and an agent design (Krichmar, 2012).

Each type of context is constituted by a set of diverse and complex elements, and the processing of all of them in artificial agents is not computationally trivial (Brooks and Mataric, 1993; Connell and Mahadevan, 1993). In this sense, this work does not pretend to be an exhaustive study of context as such. Rather, it pretends to identify and analyzed the core elements of the agent-related, environmental and task-related context to explore how they have been taken into account in CR, and then highlight the importance of the core elements interaction for behavioral flexibility under a proposed model. Here, it is suggested that, although there are innumerable elements related to the agent, the environment, and the task, the particularity of a context is constituted by means of the specific physiological needs, motivations and associated emotions that are experienced, the

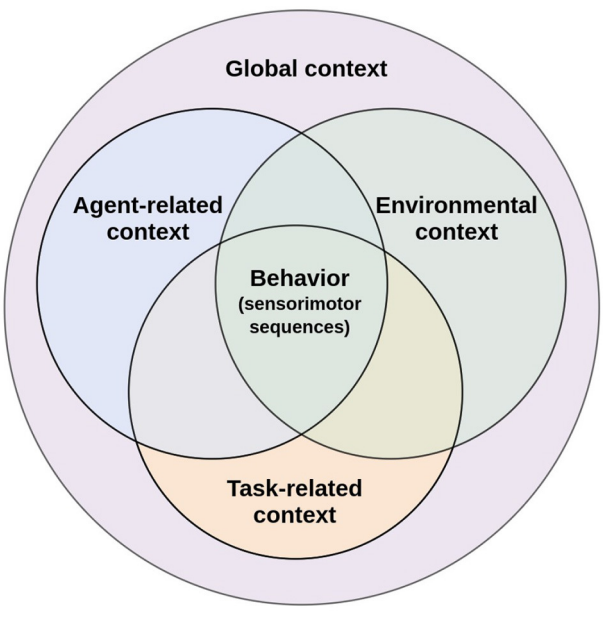



**FIGURE 1 |** Agent-related context, environmental context, and task-related context are intertwined together to influence behavior. Figure adapted from Cohen (1995).

perceived possibilities of action that a specific environment offers the agent, and the task configuration in a concrete environment.

An essential aspect of the proposed model is that it considers the monitoring of prediction error dynamics, which seems crucial for switching strategies under changing circumstances. One challenge for grounded cognition is to understand cognition in depth within the context of situated action cycles (Barsalou, 2020b). We suggest that through the monitoring of the core contextual elements, together with the monitoring of prediction error dynamics, artificial agents would autonomously select self-relevant situated tasks. We are aware that the sociocultural context plays an essential role in behavioral flexibility of social agents. However, we believe that it is essential to establish some core elements of the context associated with auto-regulation and object interaction before tackling more complex components of situated action cycles. In this way, artificial agents would enact situated action reflecting their current core context.

The structure of the paper is as follows: in Section 2, the role of the agent-related, environmental, and task-related context for behavioral flexibility is briefly explored and an overview of the processing of each one in biological agents is presented. In Sections 3–5, each type of context is addressed in more detail through their core elements and how these have been described in biological agents and then, some representative cognitive robotics implementations addressing similar elements are reviewed. In Section 6, the interaction of the three types of context in behavioral flexibility is explored through a schematic model that intertwines the core elements from each of them. Finally, Section 7 concludes the paper. For the remainder of the paper, when it reads "biological agents" it refers to living organisms, "artificial agents" refers to situated artificial robots and implementations and, when it reads "agents" it refers to both.

## 2. BEHAVIORAL FLEXIBILITY THROUGH THE LENS OF DIFFERENT TYPES OF CONTEXT

Global context includes all internal and external elements that impact and restrict the behavior of biological agents at a given moment, enticing these agents toward the performance of certain tasks or avoiding others at any given moment. Although there are countless contextual elements, they all come from three main sources: the state of the agent, the environmental conditions, and the characteristics of the task agents are engaged with in the current moment (Cohen, 1995). This allows to identify three particular types of context: agent-related, environmental, and task-related context. This section explores the role of each type of context for behavioral flexibility in biological agents. Furthermore, how each type of context is processed by the available sensory systems of these agents will be addressed. This makes it possible to establish a basis to study the notion of context within cognitive robotics in the following sections.

Flexible behavior, the ability to select the appropriate task or change strategies to adapt to the environment, is modulated by elements associated with the biological agent and the environment (Palmer et al., 2014). The elements associated with the agent that impact behavior constitute the agent-related context, which is characterized by elements such as physiological needs, emotions, as well as postural and morphological aspects. On the other hand, the environmental context relates to the characteristics of the specific environment in which the biological agent is situated, such as the spatial configuration of the objects in the environment, as well as their relational properties. Each internal or external contextual element restricts behavior to some type of task appropriate to achieve specific goals useful to the well-being of the biological agent. In this sense, behavioral flexibility is modulated by the interaction of the agent-related and environmental context. Considering both contexts, agent-related and environmental context, biological agents autonomously set goals and select appropriate tasks to achieve them according to the situation, monitoring both their needs and motivations at the current moment as well as the possibilities of action that an specific environment offers them. Task selection would be, therefore, a function of these contexts.

Once a specific task has been selected, certain elements of the biological agent and the environment become relevant to achieve the task goal, these elements constitute the task-related context (Martin et al., 2012). This type of context overlaps with agent-related and environmental context only in those elements that allow biological agents to select the appropriate sensorimotor sequence to achieve the current selected task (**Figure 1**). These elements are essential to plan and execute goal-directed movements that dynamically change during task execution, such as the situated spatial body and object configuration (perceived *via* exteroception), the body posture of the biological agent (perceived *via* proprioception), and even the area around the biological agent in which objects can be grasped and manipulated, known as peripersonal space. Every time the biological agent moves its body or an object within the task space, the task-related context is constantly "updated" to consider these changes for the planning and execution of goal-directed actions. Since its nature is a function of the selected task, this context would be redefined every time the biological agent changes tasks. Thus, the dynamics of task-related context differ from agent-related and environmental context.

From a perspective that emphasizes embodiment for the development of cognition, behavioral flexibility is achieved when it is grounded in the constant monitoring of these three contexts (**Figure 1**). This monitoring occurs through signal processing of the interoceptive, proprioceptive, and exteroceptive sensory systems. Agent-related context processing is strongly linked to interoception and proprioception. Interoception allows the perception of physiological states of the body (Schulz, 2015), which play an essential role in determining appropriate tasks for survival. Proprioception informs about body posture, the changing body position during movement, velocity, and applied force (Tuthill and Azim, 2018). Since proprioception is essential for the planning of a task, it is also closely linked to the task-related context. On the other hand, exteroception allows the processing of environmental context. Through the visual, auditory, tactile, olfactory, and gustatory sensory modalities, exteroception captures information about the changes occurring in the environmental context

(Wade, 2019). Processing environmental context helps to determine the task that better satisfies the biological agent's internal requirements according to the available external resources and the environmental configuration. Exteroception also provides information about the biological agent *situated* in the environment. Biological agents can obtain indirect information about their bodies with different exteroceptive sensory modalities, such as vision. This is essential to integrate information about the biological agent and the task being executed to guide action through perception (Kozak and Corneil, 2021). Therefore, exteroception is also associated with the task-related context.

In brief, the processing of the agent-related and environmental context provides flexibility in task selection and switching. Meanwhile, the task-related context processing provides flexibility in the planning of the sensorimotor sequence to perform a task and achieve the current goal under specific circumstances. Each contextual element constrains behavior driving the biological agent toward certain tasks and avoiding others. By endowing an artificial agent with the ability to process the current context, this agent would be able to select the task that is appropriate at a given time according to the specific circumstances at that moment. Most current artificial agents implementations, focus only on some contextual elements, those related to the task at hand, where mostly behavioral flexibility is not the object of study. The proposal here, is that, in order to achieve greater behavioral flexibility, contextual processing should be an important issue. That is why, here, some core contextual elements of each type of context have been identified that would allow artificial agents to modulate their behavior autonomously in a continuous cycle of context-sensitive actions. In Sections 3–5, we suggest core elements for the agent-related, environmental, and task-related context, respectively. At the same time, it will be addressed why these elements are considered core contextual elements for behavioral flexibility of biological agents and how they have been modeled in artificial agents.

# 3. AGENT-RELATED CONTEXT

The agent-related context refers to elements associated with the physical and physiological structure of a biological agent that modulates behavior at different hierarchical levels of organization. At a higher level, this type of context plays a fundamental role for task selection. The agent-related context allows setting specific goals, which are a priority for the biological agent to stay alive during its coupling with the environment, restricting the set of appropriate tasks possibilities to satisfy an internal need or motivation. Physiological needs, intrinsic motivation, and emotions are elements of the agent-related context that have a strong impact on this level of behavioral organization. At a lower hierarchical level, the agent-related context plays a fundamental role in the planning and execution of goal-directed and reflexive actions. Once the agent has selected a task, aspects of the agent, such as body posture and peripersonal space become relevant contextual elements for the planning and execution of the specific task. Given their role in planning and executing tasks, these contextual elements fall within the overlap of agent-related and task-related context and will be addressed as elements of task-related context. In the following, physiological needs, intrinsic motivation, and emotions will be addressed. In the first part of each subsection, the reason why said element is considered a core element of the agent-related context in biological agents will be explained. Subsequently, the second part of each subsection will provide an overview of how the addressed contextual element has been modeled in artificial agents.

## 3.1. Physiological Needs

Physiological needs, such as hunger or sleep, are sensations evoked by internal states of the biological agent that indicate a lack of nutrients, energy, or any other of the many internal conditions necessary for survival (Taormina and Gao, 2013). When physiological needs are detected by the interoceptive modality, these must be regulated to maintain the homeostasis of the biological agent (Strigo and Craig, 2016). Physiological needs are associated with motivational states that constitute action drives related to survival (Maslow, 1958). For instance, when an animal is hungry, several types of hypothalamic neurons signal this need and drive a specific task, such as foraging (Schulkin and Sterling, 2019). Thus, physiological needs are core contextual elements that have a strong impact on behavior when they are detected (Ramirez-Pedraza and Ramos, 2021). Furthermore, they modulate task activation causing an effect on the relative desirability of different tasks. In the case of hunger, this averse sensation increases the desirability of foraging and decreases the attractiveness of other tasks not associated with getting food, such as playing (Loewenstein, 2011).

Like biological agents, artificial agents must have a baseline of certain states to function properly. For example, they must have a certain level of energy, integrity in their sensors, and maintain an optimum temperature for the proper operation of their motors. In artificial agents, to keep these internal states in optimal values, some studies have focused on modeling homeostatic systems (Stradner et al., 2009; Vargas et al., 2009; Yoshida, 2017; Man and Damasio, 2019; Kelkar, 2021). Generally, artificial agents must remain in a viability zone, the set of possible states in which the operation of the system is not compromised, allowing the activation of tasks that help to regulate those internal states when they exceed a predetermined limit.

Vargas et al. (2005) proposed a model based on an artificial neural network (ANN), and on a hormone production controller. Variations in external or internal states trigger the production of a specific hormone. The level of hormones alters internal states by driving neural networks' actions through stimulation of target neurons, affecting the input weights in the ANN to perform a certain task. Once the task has been accomplished, the hormone production controller receives a negative feedback signal that ceases the production of the hormone. In another study, Moioli et al. (2009) addressed the coordination of three coupled tasks in a mobile robot: exploring the environment while avoiding obstacles, searching for a light source when fatigue is high, and searching for a black stripe in the arena when the battery is low. They use three discrete-time artificial recurrent neural networks derived from a model inspired by

gaseous modulators (Husbands et al., 1998). Each network is previously and separately evolved to accomplish a specific task. Subsequently, the output of the network is modulated by the levels of two simulated hormones associated with the levels of fatigue and hunger. The levels of hormones, together with an external stimulus, are responsible for determining the coherent coordination of behavior.

The homeostatic value of drives, together with the allostatic control for selecting appropriate behaviors to satisfy the intrinsic needs, have been modeled considering the relevance of the environmental context in Vouloutsi et al. (2013). Using a humanoid artificial agent, the designed Distributive Adaptive Control (DAC) architecture coordinated task selection depending on intrinsic drives during human-robot interaction. The DAC was based on reactive layers and adaptive layers. The reactive layers monitored the levels of the drives, sociality, exploration, survival, security, and play. The adaptive layers were responsible for the assignation of the drives' priorities, and behavior selection, depending on the current state of the world. The satisfaction of the drive and its associated homeostatic value controlled the expressed emotion of the system through facial expressions. In general, the DAC was capable of monitoring and satisfying artificial intrinsic drives, prioritizing them when several drives were competing, and organizing behavior depending on the perceived stimuli in a given environment. The DAC is a representative example of how modeling artificial internal drives and their homeostatic regulation allows an artificial agent to organize behavior autonomously responding to internal and environmental constraints.

In Kirtay et al. (2019), the authors implemented a model-free reinforcement learning (RL) framework to argue that emotion can be considered as an emergent phenomenon of a neurocomputational energy regulation mechanism in a decision-making task. This mechanism generates an internal reward signal to minimize the neural energy consumption of a sequence of actions. Each action triggers a process of visual memory recovery in which the actions to explore the environment are movements of the neck and the eyes to direct the gaze. According to the authors, the computational shortcut mechanisms on cognitive processes to facilitate energy economy give rise to emotions. In another work, Lewis and Cañamero (2016) study the role that pleasure plays in the selection of actions whether related or unrelated to the satisfaction of physiological needs. They evaluate the effects of different types of pleasures and show that pleasure, including pleasure not related to the satisfaction of physiological needs, has value for homeostatic management in terms of improved viability and greater flexibility in adaptive behavior.

A fundamental element for autonomy in artificial agents relates to energy. Most current artificial agents operate with batteries that must be replaced or recharged by the user (McFarland, 2009), so, self-charging robots would have a higher level of autonomy. In this regard, EcoBot-II is an interesting example designed to autonomously regulate its energy by converting unrefined insect biomass into useful energy using onboard microbial fuel cells with oxygen cathodes (Ieropoulos et al., 2005). The work described by Lowe et al. (2010) addresses energy-motivation autonomy where physiological information is generated by a simulated artificial metabolism as a microbial fuel cell batch. The grounding of behavior according to artificial metabolic constraints permitted the evolution of sensory anticipatory behavior in the form of simple pan/tilt active vision.

These studies show how physiological constraints impact not only sensorimotor activity but also emotional and motivational mechanisms. They allow the emergence of adaptive anticipatory behavior, prioritize tasks, and organize behavior according to the needs of artificial agents situated in a context. However, few studies address other physiological needs in artificial agents, such as engine integrity, or optimal operating temperature.

## 3.2. Emotions

There is no clear consensus about the definition of emotion, in part, because it can be defined based on its affective domain, as well as on its behavioral aspects that guide how biological agents act and respond to the environment (Soudry et al., 2011). It has been hypothesized that emotions evolved to drive behaviors that promote homeostatic processes, explaining why an emotional experience depends on the processing of interoceptive signals (Pace-Schott et al., 2019). For instance, physiological needs are strongly related to emotional experiences. Some basic emotions, such as fear, anger, disgust, sadness, happiness, and surprise could have been developed during the course of evolution and subserve adaptational strategies (Ekman, 1992, 2016).

Emotions can be generally defined as multifaceted, whole-body responses that involve coordinated changes in subjective experience, behavior, and peripheral physiology (Mauss et al., 2007). Emotions trigger responses from different biological systems, including facial expression, somatic muscle tone, tone of voice, and endocrine activity, to produce an optimal body milieu for an effective task response (Rolls, 2000). The role of these short-lived psychophysiological states encompasses coordinating behavioral response systems, shifting behavioral hierarchies, communication and social bonding, short-cut cognitive processing, facilitating storage, and recall of memories (Dolan, 2002; Phelps, 2006; Mulligan and Scherer, 2012; Tyng et al., 2017).

Emotions represent efficient modes of adaptation to changing internal and environmental demands, allowing behavioral flexibility or even triggering a task interruption when a sudden change occurs (Adolphs, 2016). They regulate behavior by associating the situation with states of positive or negative valence that express an appraisal involving a particular type of harm or benefit (Griffiths and Scarantino, 2001; Coifman and Bonanno, 2010). Thus, emotions are core contextual elements, providing direct agent-related information, regulating the selection of beneficial tasks, as well as the interruption of an ongoing task when necessary. Together, with physiological needs and intrinsic motivation, emotions drive biological agents toward behaviors that ensure their survival (Smith and Lazarus, 1990).

The computational modeling of emotions constitutes an area of growing interest in CR (Breazeal and Brooks, 2005; Ziemke and Lowe, 2009). The studies on emotions can be broadly divided into those that focus on their role in modulating behavior and those related to human-robot interaction (Arbib and Fellous, 2004). Here, we address models that highlight the role of emotions in the control of multi-task artificial agents

(Kowalczuk and Czubenko, 2010; Ghayoumi and Bansal, 2016). In these approaches, artificial agents generally learn some predefined tasks and then find their high-level coordination. Some studies associate emotions with the expected utility of each behavior. From this perspective, emotions can be considered as triggers of behavioral action sequences according to some value. The higher the value, the higher the probability of a task to be selected.

Emotions have been modeled to drive RL algorithms (Moerland et al., 2017). Gadanho and Hallam (2001) proposed a model in which emotions provided a reward value and helped a mobile robot in determining the situations in which to reevaluate decisions. The robot must maintain its energy, avoid collisions and move around a closed maze-like environment. The addressed emotions were happiness, sadness, fear, and anger. The model was implemented using a recurrent neural network in which emotions influence the perception of the state of the world. In turn, this model was integrated into an RL architecture. The intensity of emotions is associated with the internal state of the artificial agent, determined by an energy deficiency and proximity to obstacles.

Marinier and Laird (2008) implemented a cognitive architecture called *state, operator, and result* (SOAR) (Newell et al., 1987; Laird et al., 2012) as a basis for the integration of an emotion module. Emotions allow the robot to assess what stimuli attend to (sudden, relevant, pleasant), and to decide what to do with the stimulus attended. Feelings serve as a reward signal for a four-wheel-driven mobile robot. Completing a task provides the robot with a positive reward. Daglarli et al. (2009) proposed a model in which emotions and a motivational system constitute the highest control level of the architecture. The motivation module assigns behavior gain coefficients which provide an increase or decrease of the impact of the behavior. In turn, emotions determine sequences of behaviors for the planning of long-term actions according to the probabilities of transition of the emotional and behavioral states. A hidden Markov model is implemented for behavioral and emotional transition processes.

Jitviriya et al. (2015) proposed a behavioral-emotional selection model based on a self-organizing map (SOM) and a discrete stochastic state-space Markov model. The artificial agent determines the most suitable behavior and emotional expression according to internal and external situations. Firstly, the artificial agent recognizes the external situation and determines its motivation. In turn, a cognition module is used for clustering the input stimuli (the intrinsic motivation and external situation) in a SOM. Then, the robot calculates the affective and behavioral factors. The behavioral-emotional selection system is implemented with a Markov model. The basic emotions simulated in this work are normal, hope, happiness, sadness, fear, and disgust.

Emotions have also been modeled using artificial evolution. Parisi and Petrosino (2010) suggested that adding an emotional circuit to the ANN that controls behavior leads to better motivational decisions and thus greater fitness. Artificial agents must eat and drink, eat and fly away from a predator, eat and find a mating partner, eat and care for their offspring, or eat and rest to recover from physical damage. Their results show that robots with ANN that include an emotional circuit behave more effectively than robots with ANN that do not. Other approaches that use ANNs for emotional modulation of tasks focus on increasing or decreasing the synaptic efficiency of specific populations of neurons associated with tasks (Belkaid et al., 2019). In general, artificial emotions have offered an elegant approach for behavioral flexibility in artificial agents, providing a unifying way to tackle different control issues.

## 3.3. Intrinsic Motivation

Intrinsic motivation (IM) could be defined as a natural desire or interest in carrying out specific behaviors just for the pleasure and satisfaction derived while performing them, rather than for external rewards or pressures (Ryan and Deci, 2000; Sansone and Harackiewicz, 2000; Oudeyer and Kaplan, 2008; Daddaoua et al., 2016). Exploration, manipulation, curiosity, and play are considered intrinsically motivated behaviors (Ryan and Deci, 2000; Reiss, 2004; Stagnitti, 2004). White (1959) called this psychophysiological need *effectance motivation* or *mastery*. The amount of effective interaction or degree of control biological agents can have on objects, tasks, themselves, and other agents naturally motivate behavior (Deci, 1975). IM allows biological agents to acquire knowledge about themselves and their world to effectively interact with the environment, being crucial for open-ended cognitive development and for autonomy (Deci, 1975; Perry et al., 2000).

It has been observed that the most motivating situations are those with an intermediate level of novelty, this is, situations between already familiar and completely new (Berlyne, 1960). When a biological agent performs a task, an emotion with a positive or negative valence is experienced as a result of how well or bad it is performing the task. Recently, it has been suggested that the monitoring of prediction error dynamics over time is a self-regulation mechanism behind IM (Schillaci et al., 2020b). Thus, a positive emotional experience is linked to a continuous decrease in prediction error, conversely, a negative emotional experience to a continuous increase in prediction error over time (O'Reilly, 2020; Schillaci et al., 2020b). This mechanism can help to explain how biological agents select their goals, as well as why behaviors such as being curious and playful should feel good (Kiverstein et al., 2019). IM involves an ongoing cycle of finding optimal goals and interesting tasks that evoke emotions with positive valence and it is, therefore, essential for learning and encouraging interaction with the environment (Gordon, 2020; Schillaci et al., 2020b).

The tendency to be intrinsically attracted to novelty has often been used as an example of IM for guiding exploration in artificial agents (Huang and Weng, 2002; Oudeyer et al., 2007). This approach is useful to acquire optimal information gain from the novel or interesting objects to create a more accurate model of the world through curious exploration based on an intrinsic reward inversely proportional to the predictability of the environment (Schmidhuber, 1991). In knowledge-based models, the *interestingness* of an action or event derives from the comparison between the predicted sensorimotor values, based on an internal forward model, and the actual values (Oudeyer and Kaplan, 2008). The intrinsic reward for each event is

proportional to the prediction error of that event according to the learned model. Thus, interesting situations are detected by higher prediction errors.

IM allows artificial agents to autonomously select curiosity-driven goal-directed exploration behaviors and focus on goals with the optimal amount of reducible prediction errors (Schillaci et al., 2020b). Marsland et al. (2000) proposed a novelty filter using a SOM to learn representations of *normality* from sonar scans taken as a robot explores the environment. The features of the environment are clustered in the SOM. All neurons of the SOM are connected to a single output neuron. The connections to this output neuron represent the habituation process of biological neurons, recording the number of times that each winning neuron has fired. The output received from each winning neuron reduces with the number of times it fires. This allows the artificial agent to recognize novel or unusual features of the environment and forget features that repeat over time.

Competence-based models provide another measure of *interestingness*, given that it is the properties of the achievement process that will determine task selection (Oudeyer and Kaplan, 2008). Artificial agents pay little attention to those tasks that are already solved or unsolvable, for which the learning progress stays small (Colas et al., 2018). Thus, they engage in tasks associated with surprising or novel situations and can autonomously change tasks when their model has improved. The behavior is motivated by an intrinsic reward system that favors the development of competence rather than being directed to externally directed goals.

IM allows the progressive learning of more complex and hierarchically organized skills. Barto et al. (2004) proposed a strategy to explore the task space where each decision involves the execution of a temporally extended task. Agents are motivated to master tasks driven by the learning progress for each of them. Learning progress generates intrinsic rewards that determine action selection. Most implementations of IM use the RL computational framework given its inspiration in the brain reward systems (Eschmann, 2021). RL algorithms tackle the challenge of how an artificial agent can learn to approximate an optimal behavioral strategy, usually called a policy, while interacting directly with the environment. The optimality criterion of a problem is defining a reward function, an approximate solution is viewed as the skill of expertly controlling the given system (Sutton and Barto, 1998).

Luciw et al. (2011) proposed an artificial curiosity system based on RL for environmental exploration. The artificial agent builds an internal representation of its world through navigation. The reward signal is modified to contain two distinct components, one intrinsic and one external. The external component is the reward signal in classical RL, while the intrinsic reward signal is based on the measure of *interestingness* that is used as a motivational system to speed learning. The measure of *interestingness* assigns low values to patterns already known or that cannot be learned, and high values to patterns not known, but that can be discovered. The model assigns values for maximizing combined external and intrinsic rewards using a least-squares policy iteration with an internal forward model.

IM has focused on the exploration and manipulation of objects. Hart and Grupen (2012) propose that a single IM function for affordance discovery can guide long-term learning in artificial agents. Using RL, their function rewards the discovery of tasks such as finding, grasping, and placing simple objects. IM has been also used to improve the model of the artificial agent's body state and action space (Frank et al., 2014). This is achieved by guiding the exploration of states and actions using intrinsic rewards. Singh et al. (2010) consider an evolutionary perspective to define a new optimal reward framework that captures the pressure to design good primary reward functions that lead to evolutionary success across environments. They show that both intrinsic and extrinsic motivation can be understood as emergent properties of reward functions selected because they increase the fitness of learning of artificial agents across some distribution of environments. In general, IM allows learning to be more efficient by enabling the selection of novel tasks and goals with the optimal capacity for error reduction.

## 4. ENVIRONMENTAL CONTEXT

Environmental context refers to the state of the environment surrounding a biological agent at a given moment, affecting how every sensory input is processed (Nikolić, 2010). It is related to the terrain characteristics, the climate, and illumination, as well as all the entities or objects in a scene (Bloisi et al., 2016). However, the arrangement of objects is a key factor in determining the environmental context. Each scene contains specific objects that appear with a certain probability, and the spatial relations among them also present regularities (Bar, 2004). Thus, the typical spatial configuration of the environment makes it possible to distinguish different types of environmental contexts. Environmental context restricts the tasks a biological agent can select at a given moment through the action possibilities that are provided in a situation. According to Gibson (2014), affordances refer to the possibilities for action that exist by virtue of the relational properties between the environment and an agent. From a cognitive robotics' view, affordances are acquired relations through bodily interactions of an artificial agent with its environment that provide support for planning, and reside inside the artificial agent as explicit relations that enable to perceive, learn, and act (Şahin et al., 2007).

Objects by themselves do not provide action possibilities, they need to be situated in a context to stand out as relevant, affording context-dependent interactions. Each environmental context offers a field of affordances to the biological agents according to the typical objects present in it Withagen et al. (2012) and Rietveld et al. (2018). Thus, the environmental context has a predictive impact on the behavior of the biological agent, by allowing certain actions to be taken, and restricting others. Furthermore, the situated body in the environment and object configuration have predictive power in the sensorimotor sequence necessary to interact with them. Attention is deployed to process the general configuration of the objects in the environment, prioritizing those relevant regions for bodily actions (Reed and Hartley, 2021). Together, these ideas are in

line with the elements that have been suggested as necessary for physically grounding an affordance in an artificial agent. For doing so, it must be able to perform a behavior with an object given its morphology and its motor capabilities, must determine its relevance according to the artificial agent's intentions or goals, and must consider the spatio-temporal physical constraints of the objects in the environment to perform an action in the perceived context (Koppula and Saxena, 2014).

An embodied theory of spatial attention in a situated context is one that dynamically adjusts affordances of the body, the current environment, and the goals of the biological agent (Reed and Hartley, 2021). The spatial body and object configuration are fundamental elements of task-related context given their essential role in the planning and execution mechanisms for the selected task and will be addressed in Section 5. Even though many exteroceptive sensory modalities are used to obtaining environmental context information, for the sake of brevity, only visual information is addressed in this context, in both types of agents. Given the speed of contextual processing at the visual level, this sensory channel could be key to triggering predictions according to the context as stated by Bar and Aminoff (2003) and Bar (2007).

## 4.1. Spatial Configuration of the Environment

The semantic context of a scene might be extracted early enough to affect the perception of individual objects in it. Visual recognition of scenes is a fast, automatic, and reliable process (Oliva, 2005; Greene and Oliva, 2009; Lowe et al., 2018; Kaiser et al., 2019). Thorpe et al. (1996) have reported that complex natural scenes can be categorized under 150 ms. To explain this phenomenon, theories of visual perception have suggested a mode of processing based on specific spatial frequencies that would convey different information about the appearance of a stimulus (Kauffmann et al., 2015; Zhang and Li, 2019; Aghajari et al., 2020). High spatial frequencies (HSFs) represent abrupt spatial changes in visual information such as edges and correspond to configuration information and fine detail. Low spatial frequencies (LSFs) represent global information about the stimulus (Kauffmann et al., 2014). As stated by Bar and Aminoff (2003), a blurred partially analyzed image version of the visual input is projected rapidly from early visual areas toward the prefrontal cortex. LSFs in the image may provide coarse information of scenes and could reach high-order areas rapidly by conveying information through anatomical "shortcuts." HSFs, then, convey fine details of the image more slowly (Kihara and Takeda, 2010; Kauffmann et al., 2017; Petras et al., 2019).

The blurred representation of environmental context activates expectations or predictions about the most likely interpretations of the input image in higher levels, which in turn is back-projected as an initial guess to the temporal cortex to be integrated with bottom-up processing (Bar, 2007). From this perspective, a correspondence between a novel input and an existing representation similar to the input stored in memory would be activated. Then, associated representations with that similar representation would be translated into predictions. Top-down processes may facilitate recognition by limiting the number of object representations that could be considered according to the experience of the biological agent (Bar, 2004). Environmental context representation is stored in unified memory structures called context frames. Some studies have suggested that associative representations integrate information about the identity of objects and their locations (Gronau et al., 2008). These structures would bring together information about the identity of objects that are most likely to appear in a specific scene, as well as about the probable spatial relations between these objects (Bar, 2004; Gronau et al., 2008). Brady et al. (2011) argue that individual items are not represented independently of other items on the same scene. Every scene could have multiple levels of structure, from the level of feature representations to individual items to the level of ensembles of objects. Each scene representation allows simulations regarding the activated context-specific category in support of situated action (Barsalou, 2020a).

Additionally, some studies have suggested that biological agents represent knowledge about where an object is typically used in conjunction with information about how the object is used. Peelen and Caramazza (2012) provided fMRI evidence that object representations in the anterior temporal lobes would convey information about where and how an object is typically used. This favors their structural coupling with the world, generating a field of affordances relevant to each environmental context. However, it is not entirely clear how these contextual associations are stored and integrated in the brain. Once biological agents learn regularities about this coupling, fast environmental context processing would allow them to generate predictions about possible interpretations of the situation, to simulate situations, and act according to what the environmental context dictates, selecting the appropriate task in each situation taking into account also the agent-related context.

CR usually model affordances as the relation between an action, a single object, and an action effect without explicitly considering other aspects of the environmental context in which objects are embedded. Some computational algorithms for learning affordances take into account an invariant environmental context implicitly (Yukie, 2011). From an embodied perspective, this restricts the interaction with the environment and the behavioral flexibility artificial agents can acquire during the learning process. However, there exist research on environmental context can be learned through behavioral experience in artificial agents during navigation. In their pioneering work, Nolfi and Tani (1999) proposed a hierarchical architecture of prediction networks that allows a mobile artificial agent to extract spatio-temporal regularities in a a simple and structured environment in order to infer its position, as well as to detect changes in the environmental topology. In their architecture, higher layers are trained to predict the next internal state of lower layers, extracting regularities at different levels of organization. The lower-level prediction layer extracts regularities such as "walls", "corners" and "corridors", while the higher-level prediction layer, by being exposed to higher-level internal states and to shorter sequences, extracts regularities which are hidden at the sensory level, such as 'the left side wall of the large room' or "I am leaving the big room". Each prediction layer is a feedforward network with recurrent connections. After being trained in an environment consisting

of two rooms joined by a short corridor, the artificial agent is able to detect whether the corridor between the two rooms has been closed, whether a new obstacle has been placed in the environment, or whether the extension of one of the two rooms has been altered. This work is inspired by previous experiments described in Tani (1996).

In another study, Nolfi and Parisi (1996) implemented a genetic algorithm to simulate the evolution of a population of neural networks which control the behavior of mobile artificial agents that must explore efficiently an environment surrounded by walls (for a closer look at related studies see Nolfi and Floreano, 2004). In the experiments, artificial agents must be able to reach a circular target area in its environment that contains food. Since generations of artificial agents are not able to perceive the target area, they have to efficiently explore the environment to increase its chances of reaching the food arena without colliding with the walls. Each artificial agent is controlled by a feedforward neural network consisting of just an input and an output layer, without hidden units. The network includes a teaching subnetwork that determines how the standard network changes its connection weights during life. In this sense, the input generated by the teaching subnetwork can be influenced by the external context and it can teach different behaviors in different environments. Artificial agents are selected for reproduction according to their ability to explore one of the two possible environments, with dark or bright walls, respectively. Their results showed that individuals that are allowed to learn during their life perform better than those that do not learn. Although these types of studies are focused in learning environmental context through the agent's experience, these works usually pay less attention to the manipulation of objects.

On the other hand, there exist some studies that consider the environmental context to explore navigation and manipulation simultaneously (Sisbot et al., 2005). Mostly, these studies endow artificial agents with pre-set abilities so that they can perform various tasks in domestic environments. The knowledge of artificial agents usually includes databases of objects that they do not need to learn and the steps necessary to achieve goals are specified in advance. Blomqvist et al. (2020) presented a mobile manipulation system capable of perception, location, navigation, motor planning, and grasping. The artificial agent is mounted on an omnidirectional mobile base and can navigate using a 3D global pre-built map of his environment. The artificial agent builds an occupancy grid for navigation and locates itself in the environment by an online algorithm that estimates its position on the global map. During navigation, the artificial agent can detect objects through an RGB-based vision system, using a pre-trained ANN with a database of different objects. Once the task-related object is identified, the artificial agent extracts information about its position in space in order to grab it and the 3D geometry of the local scene is reconstructed in detail. Subsequently, grip pose detection algorithms are used to generate and classify a set of possible types of grasp. Finally, a path to the chosen grip position is planned and executed, the clamp is closed, and the object is retrieved from the table. The artificial agent can navigate in a laboratory, find an object on a table, take it and drop it in another place.

Asfour et al. (2006) implemented an architecture with a three-level hierarchical organization: task planning, synchronization and coordination, and execution level called sensor-actor level. Tasks are decomposed into subtasks that represent sequences of actions and contain the necessary information for execution, such as the parameters of the objects, and spatial information about the environment. The level of planning specifies the subtasks to achieve a goal and manages resources and skills. The coordination level activates actions sequentially or in parallel with the execution level. The execution level is based on control theory to execute specific control commands. This level uses specific local active models about the environment and objects. In the beginning, active models are initialized by global models, which integrate information from the environment, containing the database of objects, tasks, and abilities. The global model corresponds to long-term memory, while active models represent short-term memory.

Puigbo et al. (2015) endowed an artificial agent with predefined skills such as navigation, grasping, recognizing objects and people. They implemented the SOAR architecture as part of their approach (Newell et al., 1987; Laird et al., 2012). SOAR acts as the reasoner by selecting the actions that must be performed to achieve a goal. The control system is constituted by four main modules. Firstly, a vocal command is sent to the robot that is translated to text using an automatic speech recognition system. The semantic extractor module divides the received text into grammatical structures, from which the goal is generated. The goal is compiled in the reasoner module and sent as input to the SOAR cognitive architecture. The actions suggested by SOAR are translated as skill activations in the action nodes. The robot has information about the environment in five categories: (1) a map of the environment, (2) an ontology that contains all the actions, names of objects, people and places, (3) a database of 2D/3D models of objects that the artificial agent can recognize and grasp, (4) a database of faces that the robot can recognize and (5) a database with current knowledge of the state of the world, the artificial agent, objects and people. The information available allows the artificial agent to manipulate objects, navigate into a room, and interact with people.

Some efforts have been put into autonomous learning of the environmental context through the experience of artificial agents. However, these studies usually focus solely on environment navigation using mobile agents. Other studies have explored navigation and manipulation of objects at the same time. Generally, in these studies, environmental context is not acquired through autonomous learning. In some cases, artificial agents can plan sequences of actions. Nevertheless, the skills that they exhibit are not acquired through experience. However, it is clear that considering the environmental context extends the abilities that an artificial agent can exhibit.

## 5. TASK-RELATED CONTEXT

Biological and artificial agents interact with objects through manipulation tasks, such as grasping or pushing. Each task involves a temporarily ordered sequence of sensorimotor states

that leads to a specific goal (Grafton et al., 1998). To effectively plan and execute a sensorimotor task, agents need to acquire relevant information about themselves and the objects involved in the task. These relevant elements to achieve the task goal are determined once the task is selected and constitute the task-related context. The core elements for the planning and execution of a task suggested here are body posture, peripersonal space, and the situated body and object configuration (incoming sensory input) which dynamically change during task execution.

When grasping an object, information about its position and orientation is crucial to adapt the sensorimotor sequence accurately (Chen et al., 2014; Baltaretu et al., 2020). Given the spatial object configuration, it is possible to predict the sequence of actions that a biological agent will perform to achieve a specific goal. For instance, the type of grasp used to lift a glass would depend on whether the object is upside down or upright on a table (Rosenbaum et al., 2014). If the task involves two or more objects, the spatial relation between items becomes relevant to plan the task. Simultaneously, body posture is also essential for the execution of the sensorimotor task (Sarlegna and Sainburg, 2009). The sensorimotor sequence will also depend on the initial position of the body. This information can be directly acquired through proprioception or indirectly through incoming exteroceptive information, such as vision, which provides information about the configuration of the body situated within an environmental context.

Planning the sensorimotor sequence of a task implies that an agent has to predict the sensorimotor consequences product of its actions. During its execution, the prediction error, resulting from the difference between the predicted and the incoming sensory information, allows to dynamically adjust the sensorimotor sequence in accordance with the situated body and object configuration. Together, the body posture and object configuration would determine the sensorimotor sequence that would allow the agent to achieve the task goal (Rosenbaum et al., 2014). The body posture of an agent and its peripersonal space combined determine the location of a target relative to an extremity. The effective control of the body to avoid or manipulate objects requires an integrated neural representation of the body and the space around the body (Holmes and Spence, 2004).

## 5.1. Body Posture

Biological agents process information about the position of their limbs in space through sensory modalities, such as proprioception and vision (Sherrington, 1907; Grigg, 1994; Saunders and Knill, 2003; Saunders, 2004; Montell, 2019). The brain integrates this information in a multimodal neural representation known as body schema (Head and Holmes, 1911; Carruthers, 2008; Morasso et al., 2015; Hoffmann et al., 2020). The body schema allows to constantly monitor the body posture to trigger the planning and execution of goal-directed movements (Schillaci et al., 2016). When performing goal-directed movements, biological agents must integrate information about the body position and how this relates to extrinsic spatial coordinates of objects in the world (Sainburg et al., 2003).

Internal models have been suggested as the mechanism to code for body schema (Wolpert et al., 1995, 2001). These models allow biological agents to establish a causal relationship between their intentions and actions, as well as to anticipate the effects generated by their actions (Miall and Wolpert, 1996; Wolpert and Kawato, 1998; Kawato et al., 2003; Tanaka et al., 2020). Internal models integrate spatial body configuration and motor information to control movements and plan actions (McNamee and Wolpert, 2019). The body posture constitutes a core element of the task-related context given its determinant role in the planning and execution of action for a given task configuration (Zimmermann et al., 2012).

As infants do, artificial agents can also acquire a body schema. A common strategy is motor babbling (Demiris and Dearden, 2005; Kuniyoshi and Sangawa, 2006; Rolf et al., 2010; Houbre et al., 2021). During this process, artificial agents perform random movements which, in turn, cause changes in their sensory situation. These changes are then associated with the movements that cause them. Learning the spatio-temporal patterns that relate sensorimotor modalities with the body configuration allows artificial agents to distinguish between their own body and the environment (Diez-Valencia et al., 2019). In CR, internal models are a typical approach to allow artificial agents to acquire the sensorimotor representations necessary for prediction and action generation (Dearden and Demiris, 2005). Nevertheless, the computational tools to encode the spatial context of the body, the sensory situation, the movements as well as the approaches to map associations between them varies considerably (Schillaci et al., 2016; Nguyen et al., 2021). For example, Gama and Hoffmann (2019) study the acquisition of body schema in humanoid robots to construct map-like proprioceptive representations, resembling somatotopic representations within the brain. The joint angles of the robot are considered proprioceptive inputs and are obtained from different body configurations. Proprioceptive information serves as input to a modified SOM. The neuron activation in the maps encodes one specific joint or a combination between two or three of them as the receptive fields of neurons in the somatosensory cortex (Krubitzer et al., 2004).

Zhang et al. (2018) implemented an autoencoder to model proprioception in a humanoid robot. Interestingly, they do not consider joint angles directly as proprioceptive information, as it is typically done. Taking into account that the exact value of joint angles is unknown for biological agents, the joint configuration is the input to the network and the hidden layer is considered as proprioception. Using a multimodal variational autoencoder (VAE), Zambelli et al. (2020) proposed a system that enables an iCub to learn representations of its sensorimotor capabilities considering the spatial configuration of its body. The multimodal VAE is formed by multiple encoders and decoders, one for each sensory modality such as proprioception, vision, tactile, sound, and motor. In another study, Escobar-Juárez et al. (2016) endowed an artificial agent with the capacity of executing saccadic movements to focus a stimulus in the fovea as well as to carry out a hand-eye coordination task using multimodal representations. They proposed the Self-Organized Internal Models Architecture (SOIMA), a network of self-organized maps

interconnected with Hebbian weights. SOIMA provides coupled inverse and forward models that allow multi-modal associations of sensory and motor information.

In these studies, body schema is not adaptable as has been reported in biological agents (Iriki et al., 1996). Inspired by the flexibility of body representations, Nabeshima et al. (2006) proposed a biologically inspired model of body schema adaptation. The artificial agent reaches for and touches an object with its hand and learns to temporally integrate visual and tactile information in associative memory. If the recalled visual information is consistent with the currently obtained visual information, then the location of visual contact is considered as the location on the hand where the tactile sensation originated. If visual contact occurs not on the robot's hand, but on a given tool, then the robot is not able to adequately use the tool with the current hand trajectory controller, which induces the system to learn a new kinematic controller for the tool. In their model, the global memory is composed of two associative memories: a gating ANN to associate the visually detected target approach direction information with tactile information and, a non-monotone ANN associating tactile signals with the distance between the hand and the target. The authors suggest that tool use depends on the coherent unification of spatial and temporal aspects of multimodal information. Their model relies on the temporal integration of vision, touch and, proprioceptive information.

Learning algorithms are useful computational tools to create multimodal representations in CR, such as body schema (Hoffmann et al., 2010; Morasso and Mohan, 2021). From proprioceptive maps to multimodal representations, these studies endow artificial agents with the capacity to autonomously acquire contextual information about their own bodies. The most explored modalities in CR have been proprioception and vision. However, there is a growing interest in considering other modalities to provide artificial agents with greater behavioral flexibility (Dahiya et al., 2013; Zenha et al., 2018; Pugach et al., 2019).

## 5.2. Peripersonal Space

Peripersonal space can be understood as the reaching space of a biological agent, that is, the distance at which an object can be reached by the hand of the agent without moving the trunk (Cardinali et al., 2009; Serino, 2019). This region acts as an interface between the agent's body and the environment (Makin et al., 2008; Noel et al., 2021). Peripersonal space was also known as the flight zone and it would correspond to a margin of safety around the body (Dosey and Meisels, 1969). There is evidence about the involvement of peripersonal space in guiding involuntary defensive movements for protection. Some studies show that electrical stimulation of multimodal areas in the brain evokes a complex pattern of hand and arm movements in monkeys, similar to avoidance or defensive reactions, such as turning the head or raising the hand (Graziano et al., 2002).

Although biological agents perceive space as something continuous and unified, the processing of the peripersonal space is particularly characterized by a high degree of multi-sensory integration, mainly between visual and somatosensory (tactile and proprioceptive) information (Cardinali et al., 2009; Bertoni et al., 2020). The visually evoked responses of peripersonal multimodal neurons are modulated by the distance between the visual object and the tactile receptive field. In this way, visual information can be encoded with reference to the part of the body that contains the tactile receptive field (Cardinali et al., 2010). Such a map would give the location of the visual stimulus concerning the body surface in somatotopic coordinates. Additionally, peripersonal space includes different spatial representations, such as those around the hands and the face (Farne et al., 2005). Peripersonal space is crucial to guide movement (Graziano, 1999). It is a core contextual element of the task-related context given that it informs the body-related reachable spatial region where a specific task can be carried out.

Synthetic approaches have modeled peripersonal space centered on different parts of the body. Fuke et al. (2009) proposed a model that enables an artificial agent to acquire a head-centered peripersonal spatial representation using a SOM and Hebbian learning. Their model is inspired by the face representation in bimodal neurons found in the adjacent ventral intraparietal region of the brain, which codes the location of visual stimuli through the head-centered reference and connects visual and tactile sensations (Sereno and Huang, 2006). These neurons have been associated with the ability to avoid objects moving toward the face as a protective mechanism (Graziano and Cooke, 2006). Fuke et al. (2009) use proprioceptive information of the arm as a reference so that when the artificial agent moves his arm in front of his face the SOM is activated and learning occurs. Their simulated artificial agent learns the association of the visuo-spatial representation with the tactile representation of the face.

Juett and Kuipers (2019) recreate the learning process of peripersonal space in an artificial agent, by associating proprioceptive information of the arm and the visual perception of the hand and grippers of the agent. The peripersonal space is modeled using graphs. The nodes of the graph represent the state of the arm, and the edges correspond to safe movements. Paths represent safe trajectories from one pose to another. In their proposal, a reaching action emerges as a reliable way to hit and move an object in the environment. When an object is accidentally grasped, it moves dynamically with the hand, generating a grasping action. The learning process is modulated by a mechanism of IM and the artificial agent is capable of reaching and grasping objects based on unguided exploration.

Nguyen et al. (2019) modeled visuo-proprioceptive-tactile integration in a humanoid robot to develop reaching behaviors. They implemented a deep neural network that receives as input images from the cameras of the artificial agent and the position of the head, while the output is the arm position and tactile information of the hand and forearm. The network predicts arm configurations of successful reaching, together with information about the body part that would make contact with the objects. Finally, Jamone et al. (2012) endow an artificial agent with the ability to learn a representation of its own reachable space using motor experience. The reachable space map that they proposed uses a gaze-centered, eye-fixed reference frame. The position of a point in space can be encoded with the motor configuration

of the head and eyes of the artificial agent. Their maps are implemented using a locally weighted projection regression ANN. After learning, the artificial agent is capable of estimating the reachability of a visually detected object, even before starting the reaching movement. Together with information about the configuration of the body, peripersonal space allows artificial agents to perceive the space that surrounds them in order to carry out processes of planning and executing manipulation tasks.

## 5.3. Situated Body and Object Configuration

During task execution, it is necessary for biological agents to continuously build a visual map of the current perceived spatial body position in relation to the spatial arrangement of objects. This exteroceptive information complements the perceived body posture *via* proprioception to guide and adjust sensorimotor sequences within the peripersonal space of the biological agent. Visual working memory and attentional mechanisms are coupled by means of the action that is being executed. An action plan guides the retrieval of the appropriate sensory memory representations, and when the expected outcomes of the action are successful the representations are robustly consolidated, leading to a more rapid retrieval in the future (Olivers and Roelfsema, 2020). Thus, the content of visual working memory is to serve future behavior, in such a way that action encoding occurs in response to those visual memories of relevant objects related to the anticipated actions (Boettcher et al., 2021).

A telling example is the execution of complex grasping actions (van Polanen and Davare, 2015). The spatial information of an object interacts with the information of its physical properties to control object-oriented hand movements. This spatial object configuration must be associated with information about the body configuration in order to map spatial information about objects into body coordinates (Colby, 1998; Graziano and Gross, 1998; Bertoni et al., 2020). Thus, the situated body and object configuration is a task-related contextual element that dynamically changes during the execution of the planned sensorimotor sequences. Action plans require working memory for anticipating and chaining multiple steps, as well as the use of attentional mechanisms that are guided by the situated recurrent feedback for learning appropriate sensory-action couplings (Olivers and Roelfsema, 2020). In case of not having vision or any specific modality, it would also be expected that an integration process be carried out with those modalities available to the agent to generate predictions according to its experience. Given that all the information for planning sensorimotor sequences can not be known in advance, selective attention to relevant information during the flow of action influences subsequent action plans (Reed and Hartley, 2021). The situated action cycle has particular outcomes that potentially change the agent-related and environmental context, and these changes can also trigger further iterations of the cycle (Barsalou, 2020b).

Many studies have taken the approach of "learning by doing" to explore the consequences of self-generated actions in artificial agents. Fitzpatrick et al. (2003) showed how robots learn the effect of pushing actions on objects. In each trial, the target was placed directly in front of the robot within the task space. Then, the artificial agent executed pushing actions from any of four different initial positions. During the task, two variables were monitored, the initial proprioceptive information of the hand position and at the moment of contact and, the direction of retinal displacement of the target. In another study, Hogman et al. (2016) endow a robotic system with the ability to learn different object categories in a pushing task. The authors define categories as action-effect relations or sensorimotor contingencies, modeling the effects in an object-centered representation. The pushing task was parameterized using position and velocity. The robotic platform learns the characteristics of translation and rotation of objects and acquires knowledge with a certain degree of confidence from repeated observations of action-effect pairs. The translation is computed as the Euclidean distance between the initial and the final positions and rotation is calculated through the dihedral angle between the two planes.

Other studies have focused on addressing tool affordances. In this case, learning corresponds to finding the mapping between a set of features that describe tools and the effects that these tools produce through actions on an object. Mar et al. (2018) propose an approach where a robot learns tool affordances through interaction and generalizes them for similar tools based on their 3D geometry. During the training phase, a set of drag actions is performed by an iCub with a large number of tools grasped in different pose orientations: right, front, or left. Each trial began by placing a tool in the robot's hand. After grasping the tool, the iCub automatically detects the tool-pose it was given. Once the tool was grasped and the robot's end-effector successfully extended to the tip of the tool-pose, the robot performed a series of exploratory actions to discover the tool-poses drag affordances. Tool affordances are learned as a regression between tool-pose features and action-effect vector projections using SOMs. In this study, the initial position of the objects that were dragged is constant and object-object relations between the tool and the target object are not considered. Tool affordances are also addressed in Nabeshima et al. (2006). Interestingly, this work discusses how manipulable objects, such as tools, can become incorporated into the agent's body schema through the temporal integration of multisensory information. The contribution of Nabeshima et al. (2006) is mentioned in Section 5.1, given the emphasis their research makes on the adaptation of body schema representation.

Understanding the effects of actions is essential for planning and executing robot tasks. Paus et al. (2020) show that predicting the effects of a pushing action enables goal-oriented manipulation tasks. In this research, an artificial agent learns internal models based on objects and the spatial relations between them. The perceived scenes are represented as object-centric graphs while the internal model predicts object pose changes due to the pushing actions. The object properties are stored in the nodes of the graph while edges contain relative spatial information between object pairs. The internal model is used to predict an output graph, from which the local object position, after the push, can be extracted. This study considers the initial and final position of objects explicitly in the model and also takes into consideration spatial relation between the objects in a scene.

Using previous knowledge is crucial for performing different tasks in new situations and contexts. Khazatsky et al. (2021) developed a situated controlled system for efficient self-supervised goal-conditioned RL. A robot was trained with several previous experiences of trajectories in different tasks and contexts and tested in new environments and tasks by sampling goals from a visuomotor affordance model. After training affordances (policies), the robot was tested in new environments which contained distractor objects as well as other objects that afforded an interaction, such as opening or closing a drawer or placing an object on a pot. Importantly, these objects that afforded an interaction were not previously seen but had similar characteristics related to what they afforded (e.g., drawer with a different type of handle). In this work, learning required generalization in terms of visual affordances and their associated behaviors during online interactions to collect more data and constantly improve the associated policy. As a consequence, the policy of grasping generalizes to grasping objects and the continual learning of new tasks is faster as it benefits from increasing prior knowledge. This method of visuomotor affordance learning allows online autonomous learning of tasks in new contexts, which highlights the relevance of using prior knowledge from other contexts and their related affordances for scalable and continuous learning.

In another study, QueiSSer et al. (2021) focused on the generalization of experiences in familiar task-related contexts to those in unfamiliar task-related contexts that can be achieved through learning during vision-based goal-directed planning. In their experiments, blocks of different colors were placed at random positions in the task space, and a robot arm with a video camera was required to stack them in an arbitrary configuration specified by a visual goal. The proposed model introduces a large network composed of dynamically interacting sub-modules, which incorporates a visual working memory sub-module (VWMs), a visual attention module, and an executive network for prediction of motor states and images. This network, also controls visual attention by masks visual images in the VWM. The large network is trained by using predictive coding. Additionally, an optimal visuo-motor plan to achieve a given goal state is inferred using active inference. The experiments showed that a process of generalization occurs due to the information processing developed through the synergistic interaction between the VWM and other modules during the course of learning, in which memorizing image contents and transforming them is dissociated. After learning, the performance of the model network in generating goal directed action plans using active inference was evaluated, in cases that involved manipulating blocks with novel colors. The results showed a significant improvement in performance when using an additional VWM, compared to a case using only a single VWM. The authors suggested that the essential aspect of the mechanism acquired through learning is dissociation of visual image contents from the mechanism for their manipulation. This proposed method allows the artificial agent to flexible adapt to the new characteristic of objects during goal-directed planning.

Affordances consider the change in the task space but the representation of this change can vary drastically during task execution and within contexts. An autonomous artificial agent must be sensible to contextual changes to be able to predict the best sensorimotor sequence when performing a situated task based on the most similar previously learned situations. The use of previous experience and affordance generalization is relevant when exploring new environments. However, here we want to highlight that task-selection in a given context is also guided by the current internal needs of an agent (agent-related context), as well as by the performance expectations the agent has associated with different tasks. In biological agents, these two elements are directly linked to emotional states.

## 6. INTERACTIONIST MODEL OF CONTEXTS

The interaction of agent-related, environmental, and task-related context for behavioral flexibility is analyzed in a schematic interaction model that integrates the core contextual elements (**Figure 2**), for task selection, its execution, and disengagement when necessary. In the model, each context is perceived by its main source of sensory information. For agent-related context, interoception and proprioception are key for providing an affective and embodied context. Exteroception is central for perceiving an environmental context in a situated manner, and finally, together, proprioception, interoception, and exteroception, are fundamental for grounding a task-related context during task execution. We suggest that the model presented here is a first approximation for grounding context in artificial agents. Artificial agents will be able to manage physiological needs, and intrinsic drives for learning, considering the situated perceived environmental factors. By means of perceiving the three types of contexts and their core contextual elements, artificial agents will behave according to the changing contextual conditions. This means that artificial agents will be more prone to become competent to autonomously select tasks that are of self-relevance to 'survive', as well as tasks that promote learning, in a context-sensitive manner. This proposed interaction model is an idealized representation of the different contextual elements. In actual operation, as with other proposals (e.g., Barsalou, 2020b), one or more elements could be omitted, also, the sequence could be other than the one described here.

Biological agents learn regularities about the dynamics between the agent-related, environmental, and task-related context during their interaction with the world. It has been suggested that this association is encoded by different mechanisms, under the notion of internal models (Wolpert et al., 1995; Kawato et al., 2003; McNamee and Wolpert, 2019). Thus, biological agents learn to achieve their goals by anticipating the sensory consequences of their actions under specific contexts, and so, internal models are always context-dependent.

Internal models generate predictions about the most likely sensory consequences of self-generated actions. Biological agents always attempt to minimize the prediction error associated with predictions using two highly coupled strategies: by updating the internal model to generate better predictions or by fulfilling predictions through action to match the expected sensorimotor

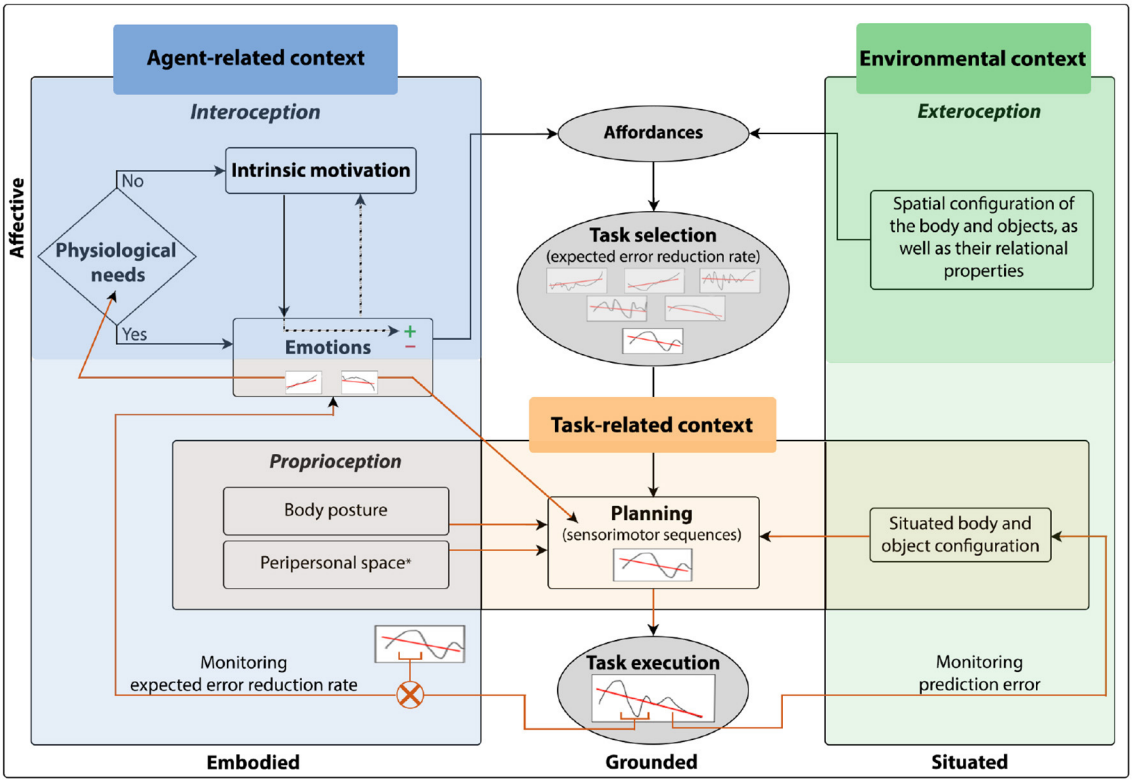



**FIGURE 2 |** Interactionist model of contexts. Schematic representation of the three different types of context and the interaction of their core elements for selection, planning, execution, and when necessary switching of a task.

states (Friston et al., 2011; Clark, 2015). Furthermore, attention has been recently drawn to the importance of the monitoring of prediction error over time when executing a task. Thus, biological agents also learn the associated rate of how prediction error is being reduced while executing a task. This rate can be understood as changes in the velocity of prediction error reduction, in such a way that it informs how well or bad a biological agent is performing a task. This monitoring of prediction error dynamics and its associated reduction expected rate is thought to play a central role in emotions and well-being (Joffily and Coricelli, 2013; Van de Cruys, 2017; Kiverstein et al., 2019; Nave et al., 2020; Hesp et al., 2021).

The positive and negative valence experienced as we act is directly related to the success of the selected behavior in reducing prediction error at the expected rate. Additionally, due that prediction error dynamics are strongly related to emotions, it has been suggested that the monitoring of the rate of error reduction can be conceived as a self-regulation mechanism for guiding behavior in artificial agents (Schillaci et al., 2020b). Thus, an artificial agent can be intrinsically motivated to autonomously select a goal associated with an optimal reducible prediction error. The capability of monitoring the error rate reduction when performing the task, allows an artificial agent to autonomously 'decide' if it should continue with the task when the pursued goal is being achieved, or if it has to be abandoned when no progress is achieved. In both scenarios, the artificial agent will be intrinsically motivated to select another goal that allows learning. It has been suggested that prediction error minimization is by itself rewarding. Decision-making based on rewards is replaced by the use of previous knowledge to avoid surprising states for survival, which is a sufficient condition to drive prediction error minimization (Friston et al., 2012).

In the model, physiological needs are central for determining which action has to be prioritized for maintaining the biological agent alive. When a physiological need is experienced, an associated emotion with a positive or negative valence, together with the environmental context, bring about the relevant affordances with which the biological agent can engage. As Rietveld et al. (2018) have suggested, biological agents respond to affordances in a context-sensitive way and affectivity is a central aspect of selective responsiveness to relevant affordances. To some extent, in the model, responding to relevant affordances for task selection and planning, can be understood as solicitations. Solicitations are those affordances that show up as relevant to a situated agent that feels immediately drawn to act a certain way (Dreyfus and Kelly, 2007). Responding with a preference to achieve a state of relative equilibrium and acting to correct for disequilibrium in relation to a dynamic field of multiple relevant affordances has been characterized as a tendency toward an optimal grip (Kiverstein et al., 2021). The best opportunities

for improving the grip with the environment come from selecting those relevant affordances that are neither too complex, nor too simple, and can potentially lead to a desired outcome of equilibrium. Here, selecting the best task among solicitations is based on their associated expected error reduction rate. This rate is learned and constantly updated during situated action cycles, being directly linked to the current competence of the agent to achieve the desired outcome (for an implementation see Schillaci et al., 2020a).

When there are no physiological needs, intrinsic motivation brings the agent to explore its environment, eliciting positive emotions related to curiosity-driven behaviors. In this situation, task selection occurs in a similar fashion, the field of relevant affordances allows the agent to select the task best suited for exploration and learning, taking into consideration its expected error reduction rate. In this regard, inspiration comes from research, on infants, understanding preferences toward optimal exploratory behaviors. In general, infants prefer to attend to stimuli that evoke an intermediate rate of complexity (Kidd et al., 2012), and to those that contain unexpected patterns of data (Stahl and Feigenson, 2015) to be able to learn based on their current competences.

Thus far, all the above mentioned, refers to the upper part of the model, the shaded areas of both agent-related and environmental context. As an example, the functioning starts on the state of the physiological needs of the agent, is there a physiological need that must be fulfilled, when yes, this evokes and emotion and together with the element in the environmental context selects a tasks from the field of affordances to fulfill the respective need. When there is no physiological need to fulfill, then intrinsic motivation is the one driving the agent to select a task in the field of affordances. For both cases, the field of relevant affordances of a particular agent is dependent on its current concerns and competences, as well as the environmental situation, also, the optimal grip on the field of affordances dynamically changes as a result of this dependency (Bruineberg and Rietveld, 2014).

In the model, once the task has been selected, either for equilibrium maintenance and self-regulation or for exploration and manipulation of the environment, the task-related context emerges. First, for planning, the proprioceptive information, framed in the task related context (both overlapping with the two other contexts), becomes relevant for the planning of sensorimotor sequences. The selected sensorimotor sequence has an expected error reduction rate, schematically shown in the planning block of the diagram as an error occurring over time and its respective slope. Then comes the execution of the selected task. During execution of the task, two types of prediction error monitoring occurs in parallel. First, the monitoring of prediction error, the predicted sensorimotor consequences of actions are compared with the actual sensorimotor input for prediction error estimation. This is shown in the task execution block, again as error over time. The perceptual fast loop occurs as the situated body and object configuration changes as the execution of the task progresses, allowing corrections when necessary. This can be though of as the fast control loop of the execution of the task, involving internal models (depicted in the overlap yellow-blue, and the overlap yellow-green, respectively). Second, there is the monitoring of the expected error reduction rate. As the task is executed, the rate of error reduction in the monitored prediction error dynamics is compared with the expected error reduction rate. In other words, the accumulated prediction error over time when executing the task allows a direct comparison between the expected error reduction rate associated to the task and the actual prediction error dynamics.

The monitoring prediction error dynamics over time and its comparison with the expected error reduction rate signal how good or bad the agent is at performing the task, or how optimal is being its grip with the environment. This comparison is schematically shown in the comparator to the right of the task execution block. The minimization of prediction error and its relation with the expected reduction rate is thought to be at the core of emotions and valence of agents actions (Kiverstein et al., 2019; Hesp et al., 2021). When a faster than expected error reduction rate occurs, produces positive emotions, motivating the agent to continue with the task. A well-done feeling, also updates the expected error reduction rate for that particular task in that particular context. This is shown by the negative slope of the error at the lower left in the emotions block, with an arrow going back down to planning and execution. A rate of minimization of the actual error which is slower than the expected one can triggers a disengagement from the task. This difference will have a negative valence and bring the system back to the slow loop by means of monitoring its current physiological needs, as well as the other core agent-related and environmental contextual elements so as to select a different task. This might also occur when the agent is not capable to minimize the error. this is shown by the error at the lower right in the emotions block, with an arrow bringing the system back to monitoring of physiological needs. When the difference between the expected error reduction rate and the actual rate is not very large, the agent might continue with the execution of the task. Still, the comparison also has an emotional valence. A positive rate of reduction is an encouragement to continue as is, whereas a negative rate might be seen as a warning or as a signal for a necessary change in the manner the task is being planned and executed (Schillaci et al., 2020b).

The model shows two different temporalities in the rate that sensory changes occur. First, a low rate of sensory changes occurs while general properties of the contexts are processed to bring relevant affordances for task selection (intense blue agent-related context; intense green environmental context). This slow loop is represented in the model by black arrows interacting with the core contextual elements for task selection and planning. Second, when the task-related context emerges, a fast rate of sensory changes occur in the environment while executing the planned sensorimotor sequence of the task (light blue, green, and yellow). This fast loop is represented in the model by orange arrows interacting with the core contextual elements during task execution. In this regard, Marchi (2020) suggested that the line that distinguishes cognition and perception can be set by considering the functional levels of the processing hierarchy. Cognitive levels, the higher levels of the hierarchy, perform more abstract and general functions to represent general knowledge

about contextual properties, and are not so susceptible to fast sensory changes that occur in the environment. On the contrary, perceptual levels, the lower levels of the hierarchy, are in close spatiotemporal proximity to sensory detectors, and are highly sensitive to fast sensory changes in the environment product of short-term actions (e.g., grasping, taking a step). Thus, the proposed model considers the sensitivity criterion proposed by Marchi (2020), in such a way that cognition is depicted by the slow loop for contextual information processing and task selection and planning, while perception is depicted by the fast loop, which is radically affected by fast sensory changes that occur during the task execution.

It is important to highlight the open question with regards to the optimal size of the time window in which prediction error dynamics has to be monitored. Different time windows of prediction error monitoring, starting from being very brief to relatively long, produce different patterns of emotional experience, as well as a different sensitivity to meaningful changes in the error reduction rate (Carver and Scheier, 1990). Recently, it has been suggested that the size of this time window should change dynamically according to 'how well or bad things are going' with respect to the expected progress (Schillaci et al., 2020a,b). Thus, when the error rate constantly decreases, meaning the agent is doing well on the task execution, the need for error monitoring diminishes. On the contrary, if prediction errors are increasing, a more careful evaluation has to be done. In computational implementations, less monitoring implies the liberation of resources. In this regard, in the proposed model, the time window by which prediction error dynamics are monitored could change dynamically based on the experienced emotions product of the differences between the expected error reduction rate and the actual reduction rate. Additionally, here it is suggested that the time window can also be influenced by the level of familiarity of the perceived environmental context. When an agent becomes familiar with a particular context, the confidence or the precision related to relevant possibilities of action increases (Friston et al., 2017a,b). Thus, in a familiar environmental context, the tasks that tend to be selected are very likely to lead to preferred outcomes (pragmatic value), and as a consequence the expected rate of error reduction is very fast. In this scenario, previous experience guides the retrieval of robustly consolidated representations for action planning that will lead to the expected outcome (Olivers and Roelfsema, 2020). Given the pragmatic value of a selected task in a familiar context, the time window by which prediction error dynamics are monitored is decreased. On the contrary, in novel or unfamiliar environmental contexts the outcomes of a set of possible tasks tend to be uncertain. Accordingly, the tasks that can be selected in a novel environmental context tend to be for exploration and learning (epistemic value). Hence, their associated expected rate of error reduction is slow. As a consequence, the time window by which prediction error dynamics are monitored is increased until more experience is gained and appropriate sensory-action couplings are consolidated.

Finally, in line with Barsalou (2020b), the interactionist model of contexts presented here offers a grounded approach to perception, cognition, and behavior. The situated action cycles in the environmental context are grounded in the task that is being executed. Central to the model is the processing of physiological needs, as well as the constant monitoring of the prediction error dynamics, which are the base for emotional states. An optimal grip with the environment is provided by the equilibrium experienced by acting in a particular situation to reduce affective tension or disequilibrium (Rietveld, 2008). Thus, a situation improves by being responsive to those relevant affordances that potentially can bring about the experience of equilibrium. Further, the proposed model highlights the particular role of the different sensory systems such as interoception, proprioception and exteroception in cognitive processes associated with the modulation of behavior. From this perspective, cognitive and perceptual processes not only occur in the brain, but are distributed in the dynamic coupling, full of affectivity, between the brain, the body, and the environment. Thus, the interactionist model of contexts is then: a) embodied in the processing of the physiological needs of agents, their morphology and their sensorimotor capabilities, b) affective, as agents act to improve the context-sensitive grip on a dynamic field of relevant affordances, c) situated in the environmental context, the current body and object configuration that, together, make the relevant affordances stand out for task selection and planning, and finally, d) grounded in the situated action cycles during task execution that trigger the processing of fast multimodal sensory changes, as well as the two types of prediction error monitoring that occurs in parallel.

## 7. DISCUSSION

Context processing plays an essential role in autonomy and behavioral flexibility of biological and artificial agents. Essentially, context is involved in all cognitive, perceptual and behavioral aspects. Endowing artificial agents with the ability to process the context in which they are situated would allow them to prioritize goals and tasks that are important for their internal self-regulation and to promote their learning and mastery of the environment. This makes context and its processing a key element for CR. The vast majority of studies in CR consider one or more contextual elements, however, the concept of context is rarely explicitly addressed. There is consensus that context acts as a set of restrictions that influence behavior, but, the discussion is open on what the notion of context actually is. Given the relevance of context not only in behavioral autonomy and flexibility but in cognition in general, this work aims to motivate the discussion about context processing within CR. In this paper, context is treated as encompassing all those elements of the agent and the environment that have an impact on decision-making and behavior. The essence of context is complex given the diverse nature of its components. Here, to address global context, a distinction has been made, analyzing context as agent-related, environmental, and task-related context. The agent-related context is characterized by elements such as physiological needs, emotions, intrinsic motivation, as well as the morphological aspects of the body. The environmental context relates to the characteristics of the specific environment in which

the agent is situated, such as the spatial configuration of the objects in the environment, as well as their relational properties. Finally, the task-related context is characterized by elements that dynamically change during the execution of the task, such as the situated spatial body and object configuration (perceived *via* exteroception), the body posture of the agent (perceived *via* proprioception), and its peripersonal space. It is suggested here, that the three types of context must be monitored at all times. When an agent is involved in the execution of a task, most of its attentional resources are devoted to achieving the goal. However, an agent can not afford to stop monitoring its physiological needs or its surroundings, big changes in any context must be attended in order to guarantee survival.

For each type of context, their core elements are analyzed separately, and several implementations in CR, representative for each core element, are described. Generally, each study focuses on different cognitive processes using a variety of mathematical and computational tools for their implementations. Here, it is proposed that establishing agent-related, environmental and task-related context allows a rapid identification of the elements considered in each study, regardless of the process modeled or computational tool used. In this sense, the classification of implementations made here, according to the core contextual elements, can shed light about the scope and limitations of the study of context in CR. At the same time, further research can be framed using this classification as a guide toward more autonomous and flexible behavior in artificial agents.

The main aim of this work is to explore and understand how the three contexts and their core elements should interact to provide behavioral flexibility in biological and artificial agents. A model is proposed integrating these core contextual elements considering their interactions and different temporalities during task selection and execution. The model gives great importance to the role of monitoring prediction error dynamics, as well as the expected error reduction rate. The agent-related context, together with the environmental context bring about a field of affordances at a given moment. Task selection is made on the field of relevant affordances according to the expected prediction error reduction rate for each task. Monitoring of prediction error dynamics allows online corrections of the planned sensorimotor sequence, by comparing predictions with incoming sensory information. All these, occur in the grounded task-related context during the agent's situated action cycles. Monitoring prediction error over time, as the task is executed, and comparing it with the expected prediction error reduction rate allows an agent to be sensible to its performance. This sensitivity signals if it is appropriate to continue execution, when results are positive and it "feels good," or autonomously switch task, when things occur not as expected, and the task becomes "frustrating." The model also includes two temporal resolutions, a slower one for cognition and a faster one for perception and situated action cycles.

Finally, the interactionist model of contexts suggested here is embodied, affective, and situated, by means of the monitoring of the agent-related and environmental core contextual elements. Additionally, it is grounded in the processing of the task-related context and the associated situated action cycles during task execution. The model suggests how artificial agents should monitor the core contextual elements of the agent-related and environmental context to give rise to the task-related context based on the field of relevant affordances, their associated expected error reduction rate and its positive or negative emotional valence, reflecting a tendency toward an optimal grip. This capability allows agents to autonomously select a task, its planning, execution, and monitoring for behavioral flexibility. In this regard, the model could shed light on the complexity of the dynamics of affordances' activation and to what extent the context filters this activation (see Borghi, 2018, for an extensive analysis of this issue). The modeling of context is essential to study the structural coupling between agents and their environment. The model presented here aims to contribute in this direction, as well as in clarifying the notion of context for behavioral flexibility, not only in artificial agents, also in biological agents.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

DV contributed to the conception of the study and drafted the first manuscript. AC, GS, and BL guided and contributed to the writing of the article. All authors contributed to the conceptual design of the work and to the development of the proposed model. Finally, all authors contributed to manuscript revision and approved the submitted version.

## FUNDING

This research was supported by the Consejo Nacional de Ciencia y Tecnología (CONACyT; grant no. 517333).

## REFERENCES

Adnan Mohsin Abdulazeez, F. S. F. (2021). Vision-based mobile robot controllers: a scientific review. *Turk. J. Comput. Math. Educ.* 12, 1563–1580. doi: 10.17762/turcomat.v12i6.2695

Adolphs, R. (2016). How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Soc. Cogn. Affect. Neurosci.* 12, 24–31. doi: 10.1093/scan/nsw153

Aghajari, S., Vinke, L. N., and Ling, S. (2020). Population spatial frequency tuning in human early visual cortex. *J. Neurophysiol.* 123, 773–785. doi: 10.1152/jn.00291.2019

Arbib, M. A., and Fellous, J.-M. (2004). Emotions: from brain to robot. *Trends Cogn. Sci.* 8, 554–561. doi: 10.1016/j.tics.2004.10.004

Asada, M. (2020). Rethinking autonomy of humans and robots. *J. Artif. Intell. Consciousness* 07, 141–153. doi: 10.1142/S2705078520500083

Asada, M., MacDorman, K. F., Ishiguro, H., and Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Rob. Auton. Syst*. 37, 185–193. doi: 10.1016/S0921-8890(01)00157-9

Asfour, T., Regenstein, K., Azad, P., Schroder, J., Bierbaum, A., Vahrenkamp, N., et al. (2006). "ARMAR-III: an integrated humanoid platform for sensory-motor control," in *2006 6th IEEE-RAS International Conference on Humanoid Robots* (Genova: IEEE), 169–175.

Baltaretu, B. R., Monaco, S., Velji-Ibrahim, J., Luabeya, G. N., and Crawford, J. D. (2020). Parietal cortex integrates saccade and object orientation signals to update grasp plans. *J. Neurosci*. 40, 4525–4535. doi: 10.1523/JNEUROSCI.0300-20.2020

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci*. 5, 617–629. doi: 10.1038/nrn1476

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends Cogn. Sci*. 11, 280–289. doi: 10.1016/j.tics.2007.05.005

Bar, M., and Aminoff, E. (2003). Cortical analysis of visual context. *Neuron* 38, 347–358. doi: 10.1016/S0896-6273(03)00167-3

Barsalou, L. W. (2020a). "Categories at the interface of cognition and action," in *Building Categories in Interaction: Linguistic Resources at Work*, eds I. Fiorentini, E. Goria, and C. Mauri (Amsterdam: John Benjamins), 35–72.

Barsalou, L. W. (2020b). Challenges and opportunities for grounding cognition. *J. Cogn*. 3, 116. doi: 10.5334/joc.116

Barto, A. G., Singh, S., and Chentanez, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of International Conference on Developmental Learning (ICDL)* (Cambridge, MA: MIT Press), 112–119.

Bazire, M., and Brézillon, P. (2005). "Understanding context before using it," in *Modeling and Using Context* (Berlin; Heidelberg: Springer Berlin Heidelberg), 29–40.

Belkaid, M., Cuperlier, N., and Gaussier, P. (2019). Autonomous cognitive robots need emotional modulations: Introducing the eMODUL model. *IEEE Trans. Syst. Man Cybern. Syst*. 49, 206–215. doi: 10.1109/TSMC.2018.2792542

Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity*. New York, NY: McGraw-Hill Book Company.

Bertoni, T., Magosso, E., and Serino, A. (2020). From statistical regularities in multisensory inputs to peripersonal space representation and body ownership: Insights from a neural network model. *Eur. J. Neurosci*. 53, 611–636. doi: 10.1111/ejn.14981

Bloisi, D. D., Nardi, D., Riccio, F., and Trapani, F. (2016). "Context in robotics and information fusion," in *Context-Enhanced Information Fusion* (Cham: Springer International Publishing), 675–699.

Blomqvist, K., Breyer, M., Cramariuc, A., Förster, J., Grinvald, M., Tschopp, F., et al. (2020). Go fetch: mobile manipulation in unstructured environments. *arXiv preprint* arXiv:2004.00899. doi: 10.48550/arXiv.2004.00899

Boettcher, S. E., Gresch, D., Nobre, A. C., and van Ede, F. (2021). Output planning at the input stage in visual working memory. *Sci. Adv*. 7, eabe8212. doi: 10.1126/sciadv.abe8212

Borghi, A. M. (2018). Affordances, context and sociality. *Synthese* 199, 12485–12515. doi: 10.1007/s11229-018-02044-1

Brady, T. F., Konkle, T., and Alvarez, G. A. (2011). A review of visual memory capacity: beyond individual items and toward structured representations. *J. Vis*. 11, 4–4. doi: 10.1167/11.5.4

Breazeal, C., and Brooks, R. (2005). "Robot emotion," in *Who Needs Emotions*? (Oxford: Oxford University Press), 271–310.

Brooks, R. A., and Mataric, M. J. (1993). "Real robots, real learning problems," in *Robot Learning* (New York, NY: Springer US), 193–213.

Bruineberg, J., and Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Front. Hum. Neurosci*. 8, 599. doi: 10.3389/fnhum.2014.00599

Cangelosi, A., Bongard, J., Fischer, M. H., and Nolfi, S. (2015). "Embodied intelligence," in *Springer Handbook of Computational Intelligence* (Berlin; Heidelberg: Springer Berlin Heidelberg), 697–714.

Cangelosi, A., and Schlesinger, M. (2015). *Developmental Robotics:From Babies to Robots*. Berlin; Heidelberg: The MIT Press.

Cardinali, L., Brozzoli, C., and Farnè, A. (2009). Peripersonal space and body schema: two labels for the same concept? *Brain Topogr*. 21, 252–260. doi: 10.1007/s10548-009-0092-7

Cardinali, L., Brozzoli, C., and Farnè, A. (2010). "Peripersonal space and body schema," in *Encyclopedia of Behavioral Neuroscience* (Amsterdam: Elsevier), 40–46.

Carruthers, G. (2008). Types of body representation and the sense of embodiment. *Conscious Cogn*. 17, 1302–1316. doi: 10.1016/j.concog.2008.02.001

Carver, C. S., and Scheier, M. F. (1990). Origins and functions of positive and negative affect: a control-process view. *Psychol. Rev*. 97, 19. doi: 10.1037/0033-295X.97.1.19

Chen, Y., Monaco, S., Byrne, P., Yan, X., Henriques, D. Y. P., and Crawford, J. D. (2014). Allocentric versus egocentric representation of remembered reach targets in human cortex. *J. Neurosci*. 34, 12515–12526. doi: 10.1523/JNEUROSCI.1445-14.2014

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Cohen, P. R. (1995). *Empirical Methods for Artificial Intelligence, Vol. 139*. Cambridge, MA: MIT Press.

Coifman, K. G., and Bonanno, G. A. (2010). When distress does not become depression: emotion context sensitivity and adjustment to bereavement. *J. Abnorm. Psychol*. 119, 479–490. doi: 10.1037/a0020113

Colas, C., Fournier, P., Sigaud, O., Chetouani, M., and Oudeyer, P.-Y. (2018). Curious: Intrinsically motivated modular multi-goal reinforcement learning. *arXiv e-prints*, arXiv-1810. doi: 10.48550/arXiv.1810.06284

Colby, C. L. (1998). Action-oriented spatial reference frames in cortex. *Neuron*. 20, 15–24. doi: 10.1016/S0896-6273(00)80429-8

Connell, J. H. and Mahadevan, S. (Eds.). (1993). *Robot Learning*. New York, NY: Springer US.

Daddaoua, N., Lopes, M., and Gottlieb, J. (2016). Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates. *Sci. Rep*. 6, 20202. doi: 10.1038/srep20202

Daglarli, E., Temeltas, H., and Yesiloglu, M. (2009). Behavioral task processing for cognitive robots using artificial emotions. *Neurocomputing* 72, 2835–2844. doi: 10.1016/j.neucom.2008.07.018

Dahiya, R. S., Mittendorfer, P., Valle, M., Cheng, G., and Lumelsky, V. J. (2013). Directions toward effective utilization of tactile skin: a review. *IEEE Sens. J*. 13, 4121–4138. doi: 10.1109/JSEN.2013.2279056

Dearden, A., and Demiris, Y. (2005). "Learning forward models for robots," in *IJCAI, Vol* (London). 5, 1440.

Deci, E. L. (1975). "The intrinsic motivation of behavior," in *Intrinsic Motivation* (New York, NY; London: Springer US), 93–125.

Demiris, Y., and Dearden, A. (2005). "From motor babbling to hierarchical learning by imitation: a robot developmental pathway," in *Proceedings of the 5th International Workshop on Epigenetic Robotics Systems* (Lund: Lund University), 31-37. Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.346&rep=rep1&type=pdf

Diez-Valencia, G., Ohashi, T., Lanillos, P., and Cheng, G. (2019). Sensorimotor learning for artificial body perception. *arXiv preprint* arXiv:1901.09792. doi: 10.48550/arXiv.1901.09792

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358

Dosey, M. A., and Meisels, M. (1969). Personal space and self-protection. *J. Pers. Soc. Psychol*. 11, 93–97. doi: 10.1037/h0027040

Dreyfus, H., and Kelly, S. D. (2007). Heterophenomenology: Heavy-handed sleight-of-hand. *Phenomenol. Cogn. Sci*. 6, 45–55. doi: 10.1007/s11097-006-9042-y

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot*. 6, 169–200. doi: 10.1080/02699939208411068

Ekman, P. (2016). What scientists who study emotion agree about. *Perspect. Psychol. Sci*. 11, 31–34. doi: 10.1177/1745691615596992

Eschmann, J. (2021). "Reward function design in reinforcement learning," in *Reinforcement Learning Algorithms: Analysis and Applications* (Cham: Springer International Publishing), 25–33.

Escobar-Juárez, E., Schillaci, G., Hermosillo-Valadez, J., and Lara-Guzmán, B. (2016). A self-organized internal models architecture for coding sensory-motor schemes. *Front. Rob. AI* 3, 22. doi: 10.3389/frobt.2016.00022

Farne, A., Dematte, M. L., and Ladavas, E. (2005). Neuropsychological evidence of modular organization of the near peripersonal space. *Neurology* 65, 1754–1758. doi: 10.1212/01.wnl.0000187121.30480.09

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). "Learning about objects through action-initial steps towards artificial cognition, in *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, *Vol. 3* (Taipei: IEEE), 3140–3145.

Frank, M., Leitner, J., Stollenga, M., Förster, A., and Schmidhuber, J. (2014). Curiosity driven reinforcement learning for motion planning on humanoids. *Front. Neurorob.* 7, 25. doi: 10.3389/fnbot.2013.00025

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Mattout, J., and Kilner, J. (2011). Action understanding and active inference. *Biol. Cybern.* 104, 137–160. doi: 10.1007/s00422-011-0424-z

Friston, K., Samothrakis, S., and Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biol. Cybern.* 106, 523–541. doi: 10.1007/s00422-012-0512-8

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Fuke, S., Ogino, M., and Asada, M. (2009). Acquisition of the head-centered peripersonal spatial representation found in VIP neuron. *IEEE Trans. Auton. Ment. Dev.* 1, 131–140. doi: 10.1109/TAMD.2009.2031013

Gadanho, S. C., and Hallam, J. (2001). Robot learning driven by emotions. *Adapt. Behav.* 9, 42–64. doi: 10.1177/105971230200900102

Gama, F., and Hoffmann, M. (2019). The homunculus for proprioception: Toward learning the representation of a humanoid robot's joint space using self-organizing maps. *arXiv preprint* arXiv:1909.02295. doi: 10.48550/arXiv.1909.02295

Ghayoumi, M., and Bansal, A. K. (2016). "Multimodal architecture for emotion in robots using deep learning," in *2016 Future Technologies Conference (FTC)* (San Francisco, CA: IEEE), 901–907.

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception*. New York, NY: Psychology Press.

Gordon, G. (2020). Infant-inspired intrinsically motivated curious robots. *Curr. Opin. Behav. Sci.* 35, 28–34. doi: 10.1016/j.cobeha.2020.05.010

Grafton, S. T., Hazeltine, E., and Ivry, R. B. (1998). Abstract and effector-specific representations of motor sequences identified with PET. *J. Neurosci.* 18, 9420–9428. doi: 10.1523/JNEUROSCI.18-22-09420.1998

Graziano, M. S., and Cooke, D. F. (2006). Parieto-frontal interactions, personal space, and defensive behavior. *Neuropsychologia* 44, 845–859. doi: 10.1016/j.neuropsychologia.2005.09.009

Graziano, M. S., and Gross, C. G. (1998). Spatial maps for the control of movement. *Curr. Opin. Neurobiol.* 8, 195–201. doi: 10.1016/S0959-4388(98)80140-2

Graziano, M. S., Taylor, C. S., and Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron* 34, 841–851. doi: 10.1016/S0896-6273(02)00698-0

Graziano, M. S. A. (1999). Where is my arm? the relative role of vision and proprioception in the neuronal representation of limb position. *Proc. Natl. Acad. Sci. U.S.A.* 96, 10418–10421. doi: 10.1073/pnas.96.18.10418

Greene, M. R., and Oliva, A. (2009). The briefest of glances. *Psychol. Sci.* 20, 464–472. doi: 10.1111/j.1467-9280.2009.02316.x

Griffiths, P., and Scarantino, A. (2001). "Emotions in the wild," in *The Cambridge Handbook of Situated Cognition* (Cambridge: Cambridge University Press), 437–453.

Grigg, P. (1994). Peripheral neural mechanisms in proprioception. *J. Sport Rehabil.* 3, 2–17. doi: 10.1123/jsr.3.1.2

Gronau, N., Neta, M., and Bar, M. (2008). Integrated contextual representation for objects' identities and their locations. *J. Cogn. Neurosci.* 20, 371–388. doi: 10.1162/jocn.2008.20027

Hart, S., and Grupen, R. (2012). "Intrinsically motivated affordance discovery and modeling," in *Intrinsically Motivated Learning in Natural and Artificial Systems* (Berlin; Heidelberg: Springer Berlin Heidelberg), 279–300.

Head, H., and Holmes, G. (1911). Sensory disturbances from cerebral lesions. *Brain* 34, 102–254. doi: 10.1093/brain/34.2-3.102

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Comput.* 33, 398–446. doi: 10.1162/neco_a_01341

Hoffmann, M., Marques, H., Arieta, A., Sumioka, H., Lungarella, M., and Pfeifer, R. (2010). Body schema in robotics: a review. *IEEE Trans. Auton. Ment. Dev.* 2, 304–324. doi: 10.1109/TAMD.2010.2086454

Hoffmann, M., Pitti, A., Jamone, L., Somogyi, E., and Lanillos, P. (Eds.). (2020). *Body Representations, Peripersonal Space, and the Self: Humans, Animals, Robots*. Lausanne: Frontiers Media SA.

Hogman, V., Bjorkman, M., Maki, A., and Kragic, D. (2016). A sensorimotor learning framework for object categorization. *IEEE Trans. Cogn. Dev. Syst.* 8, 15–25. doi: 10.1109/TAMD.2015.2463728

Holmes, N. P., and Spence, C. (2004). The body schema and multisensory representation(s) of peripersonal space. *Cogn. Process.* 5, 94–105. doi: 10.1007/s10339-004-0013-3

Houbre, Q., Angleraud, A., and Pieters, R. (2021). "Balancing exploration and exploitation: a neurally inspired mechanism to learn sensorimotor contingencies," in *Springer Proceedings in Advanced Robotics* (Cham: Springer International Publishing), 59–73.

Huang, X., and Weng, J. (2002). "Novelty and reinforcement learning in the value system of developmental robots," in *Proceedings of the 2nd International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 2002* (Lund), 47–55.

Husbands, P., Smith, T., Jakobi, N., and O'Shea, M. (1998). Better living through chemistry: Evolving GasNets for robot control. *Conn. Sci.* 10, 185–210. doi: 10.1080/095400998116404

Ieropoulos, I., Melhuish, C., Greenman, J., and Horsfield, I. (2005). EcoBot-II: An artificial agent with a natural metabolism. *Int. J. Adv. Rob. Syst.* 2, 31. doi: 10.5772/5777

Jamone, L., Natale, L., Hashimoto, K., Sandini, G., and Takanishi, A. (2012). "Learning the reachable space of a humanoid robot: a bio-inspired approach," in *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)* (Rome: IEEE), 1148–1154.

Jitviriya, W., Koike, M., and Hayashi, E. (2015). "Behavior selection system based on emotional variations," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Kobe: IEEE), 462–467.

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput. Biol.* 9, e1003094. doi: 10.1371/journal.pcbi.1003094

Juett, J., and Kuipers, B. (2019). Learning and acting in peripersonal space: Moving, reaching, and grasping. *Front. Neurorobot.* 13, 4. doi: 10.3389/fnbot.2019.00004

Kaiser, D., Häberle, G., and Cichy, R. M. (2019). Cortical sensitivity to natural scene structure. *Hum. Brain Mapp.* 41, 1286–1295. doi: 10.1002/hbm.24875

Kauffmann, L., Chauvin, A., Guyader, N., and Peyrin, C. (2015). Rapid scene categorization: role of spatial frequency order, accumulation mode and luminance contrast. *Vision Res.* 107, 49–57. doi: 10.1016/j.visres.2014.11.013

Kauffmann, L., Ramanoël, S., and Peyrin, C. (2014). The neural bases of spatial frequency processing during scene perception. *Front. Integr. Neurosci.* 8, 37. doi: 10.3389/fnint.2014.00037

Kauffmann, L., Roux-Sibilon, A., Beffara, B., Mermillod, M., Guyader, N., and Peyrin, C. (2017). How does information from low and high spatial frequencies interact during scene categorization? *Vis. Cogn.* 25, 853–867. doi: 10.1080/13506285.2017.1347590

Kawato, M., Kuroda, T., Imamizu, H., Nakano, E., Miyauchi, S., and Yoshioka, T. (2003). "Internal forward models in the cerebellum: fMRI study on grip force and load force coupling," in *Progress in Brain Research* (Amsterdam: Elsevier), 171–188.

Kelkar, A. (2021). Cognitive homeostatic agents. *arXiv preprint* arXiv:2103.03359. doi: 10.5555/3461017.3461021

Khazatsky, A., Nair, A., Jing, D., and Levine, S. (2021). What can i do here? learning new skills by imagining visual affordances. *arXiv preprint* arXiv:2106.00671. doi: 10.1109/ICRA48506.2021.9561692

Kidd, C., Piantadosi, S. T., and Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE* 7, e36399. doi: 10.1371/journal.pone.0036399

Kihara, K., and Takeda, Y. (2010). Time course of the integration of spatial frequency-based information in natural scenes. *Vision Res.* 50, 2158–2162. doi: 10.1016/j.visres.2010.08.012

Kirtay, M., Vannucci, L., Albanese, U., Laschi, C., Oztop, E., and Falotico, E. (2019). Emotion as an emergent phenomenon of the neurocomputational energy regulation mechanism of a cognitive agent in a decision-making task. *Adaptive Behav*. 29, 55–74. doi: 10.1177/1059712319880649

Kiverstein, J., Miller, M., and Rietveld, E. (2019). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese* 196, 2847–2869. doi: 10.1007/s11229-017-1583-9

Kiverstein, J., Van Dijk, L., and Rietveld, E. (2021). The field and landscape of affordances: Koffka's two environments revisited. *Synthese* 198, 2279–2296. doi: 10.1007/s11229-019-02123-x

Koppula, H. S., and Saxena, A. (2014). "Physically grounded spatio-temporal object affordances," in *European Conference on Computer Vision* (Cham; Heidelberg; New York, NY; Dordrecht; London: Springer), 831–847.

Kowalczuk, Z., and Czubenko, M. (2010). "Model of human psychology for controlling autonomous robots," in *2010 15th International Conference on Methods and Models in Automation and Robotics* (Miedzyzdroje: IEEE), 31–36.

Kozak, R. A., and Corneil, B. D. (2021). High contrast, moving targets in an emerging target paradigm promote fast visuomotor responses during visually guided reaching. *bioRxiv*. doi: 10.1152/jn.00057.2021

Krichmar, J. L. (2012). Design principles for biologically inspired cognitive robotics. *Biol. Inspired Cogn. Architect*. 1, 73–81. doi: 10.1016/j.bica.2012.04.003

Krubitzer, L., Huffman, K. J., Disbrow, E., and Recanzone, G. (2004). Organization of area 3a in macaque monkeys: contributions to the cortical phenotype. *J. Comp. Neurol*. 471, 97–111. doi: 10.1002/cne.20025

Kuniyoshi, Y., and Sangawa, S. (2006). Early motor development from partially ordered neural-body dynamics: experiments with a cortico-spinal-musculo-skeletal model. *Biol. Cybern*. 95, 589–605. doi: 10.1007/s00422-006-0127-z

Laird, J. E., Kinkade, K. R., Mohan, S., and Xu, J. Z. (2012). "Cognitive robotics using the soar cognitive architecture," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, ON: AAAI), 46–54.

Lara, B., Astorga, D., Mendoza-Bock, E., Pardo, M., Escobar, E., and Ciria, A. (2018). Embodied cognitive robotics and the learning of sensorimotor schemes. *Adapt. Behav*. 26, 225–238. doi: 10.1177/1059712318780679

Lea, S. E. G., Chow, P. K. Y., Leaver, L. A., and McLaren, I. P. L. (2020). Behavioral flexibility: a review, a model, and some exploratory tests. *Learn. Behav*. 48, 173–187. doi: 10.3758/s13420-020-00421-w

Lewis, M., and Cañamero L. (2016). Hedonic quality or reward? a study of basic pleasure in homeostasis and decision making of a motivated autonomous robot. *Adapt. Behav*. 24, 267–291. doi: 10.1177/1059712316666331

Loewenstein, G. (2011). "Out of control," in *Advances in Behavioral Economics* (Princeton, NJ: Princeton University Press), 689–724.

Lowe, M. X., Rajsic, J., Ferber, S., and Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *Cortex* 106, 275–287. doi: 10.1016/j.cortex.2018.06.006

Lowe, R., Montebelli, A., Ieropoulos, I., Greenman, J., Melhuish, C., and Ziemke, T. (2010). "Grounding motivation in energy autonomy: a study of artificial metabolism constrained robot dynamics," in *Artificial Life XII: Proceedings of the 12th International Conference on the Synthesis and Simulation of Living Systems, ALIFE 2010* (Odense: MIT Press), 725–732.

lriki, A., Tanaka, M., and Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurones. *Neuroreport* 7, 2325–2330. doi: 10.1097/00001756-199610020-00010

Luciw, M., Graziano, V., Ring, M., and Schmidhuber, J. (2011). "Artificial curiosity with planning for autonomous perceptual and cognitive development," in *2011 IEEE International Conference on Development and Learning (ICDL)* (Frankfurt am Main: IEEE), 1–8.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Conn. Sci*. 15, 151–190. doi: 10.1080/09540090310001655110

Makin, T. R., Holmes, N. P., and Ehrsson, H. H. (2008). On the other hand: Dummy hands and peripersonal space. *Behav. Brain Res*. 191, 1–10. doi: 10.1016/j.bbr.2008.02.041

Man, K., and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell*. 1, 446–452. doi: 10.1038/s42256-019-0103-7

Mar, T., Tikhanoff, V., and Natale, L. (2018). What can i do with this tool? self-supervised learning of tool affordances from their 3-d geometry. *IEEE Trans. Cogn. Dev. Syst*. 10, 595–610. doi: 10.1109/TCDS.2017.2717041

Marchi, F. (2020). "The cognitive processing hierarchy," in *The Attentional Shaping of Perceptual Experience* (Cham: Springer), 17–32.

Marinier, R. P., and Laird, J. E. (2008). "Emotion-driven reinforcement learning," in *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 30*, 115–120.

Marsland, S., Nehmzow, U., and Shapiro, J. (2000). A real-time novelty detector for a mobile robot. *arXiv preprint* cs/0006006. doi: 10.48550/arXiv.cs/0006006

Martin, J., van Moorter, B., Revilla, E., Blanchard, P., Dray, S., Quenette, P.-Y., et al. (2012). Reciprocal modulation of internal and external factors determines individual movements. *J. Anim. Ecol*. 82, 290–300. doi: 10.1111/j.1365-2656.2012.02038.x

Maslow, A. H. (1958). "A dynamic theory of human motivation," in *Understanding Human Motivation* (Howard Allen Publishers), 26–47.

Mauss, I. B., Bunge, S. A., and Gross, J. J. (2007). Automatic emotion regulation. *Soc. Pers. Psychol. Compass*. 1, 146–167. doi: 10.1111/j.1751-9004.2007.00005.x

McFarland, D. (2009). Guilty robots, happy dogs: the question of alien minds. *Choice Rev. Online* 46, 46-2662–46-2662. doi: 10.5860/CHOICE.46-2662

McNamee, D., and Wolpert, D. M. (2019). Internal models in biological control. *Ann. Rev. Control Rob. Auton. Syst*. 2, 339–364. doi: 10.1146/annurev-control-060117-105206

Miall, R., and Wolpert, D. (1996). Forward models for physiological motor control. *Neural Netw*. 9, 1265–1279. doi: 10.1016/S0893-6080(96)00035-4

Min, H., Yi, C., Luo, R., Zhu, J., and Bi, S. (2016). Affordance research in developmental robotics: a survey. *IEEE Trans. Cogn. Dev. Syst*. 8, 237–255. doi: 10.1109/TCDS.2016.2614992

Moerland, T. M., Broekens, J., and Jonker, C. M. (2017). Emotion in reinforcement learning agents and robots: a survey. *Mach. Learn*. 107, 443–480. doi: 10.1007/s10994-017-5666-0

Mohan, V., Morasso, P., Sandini, G., and Kasderidis, S. (2013). Inference through embodied simulation in cognitive robots. *Cogn. Comput*. 5, 355–382. doi: 10.1007/s12559-013-9205-4

Moioli, R. C., Vargas, P. A., and Husbands, P. (2009). "A multiple hormone approach to the homeostatic control of conflicting behaviours in an autonomous mobile robot," in *2009 IEEE Congress on Evolutionary Computation* (Trondheim: IEEE), 47–54.

Montell, C. (2019). Coordinated movement: watching proprioception unfold. *Curr. Biol*. 29, R202-R205. doi: 10.1016/j.cub.2019.02.004

Morasso, P., Casadio, M., Mohan, V., Rea, F., and Zenzeri, J. (2015). Revisiting the body-schema concept in the context of whole-body postural-focal dynamics. *Front. Hum. Neurosci*. 9, 83. doi: 10.3389/fnhum.2015.00083

Morasso, P., and Mohan, V. (2021). The body schema: neural simulation for covert and overt actions of embodied cognitive agents. *Curr. Opin. Physiol*. 19, 219–225. doi: 10.1016/j.cophys.2020.11.009

Mulligan, K., and Scherer, K. R. (2012). Toward a working definition of emotion. *Emot. Rev*. 4, 345–357. doi: 10.1177/1754073912445818

Nabeshima, C., Kuniyoshi, Y., and Lungarella, M. (2006). Adaptive body schema for robotic tool-use. *Adv. Rob*. 20, 1105–1126. doi: 10.1163/156855306778522550

Nave, K., Deane, G., Miller, M., and Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisc. Rev. Cogn. Sci*. 11, e1542. doi: 10.1002/wcs.1542

Newell, A., Rosenbloom, P. S., and Laird, J. E. (1987). SOAR: *An architecture for general intelligence*. Technical report, Stanford University CA Department of Computer Science.

Nguyen, P. D., Hoffmann, M., Pattacini, U., and Metta, G. (2019). "Reaching development through visuo-proprioceptive-tactile integration on a humanoid robot - a deep learning approach," in *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Oslo: IEEE), 163–170.

Nguyen, P. D. H., Georgie, Y. K., Kayhan, E., Eppe, M., Hafner, V. V., and Wermter, S. (2021). Sensorimotor representation learning for an "active self" in robots: a model survey. *Künstliche Intell*. 35, 9–35. doi: 10.1007/s13218-021-00703-z

Nikolić, D. (2010). The brain is a context machine. *Rev. Psychol*. 17, 33–38.

Noel, J.-P., Bertoni, T., and Serino, A. (2021). "Peri-personal space as an interface for self-environment interaction," in *The World at Our Fingertips* (Oxford: Oxford University Press), 17–46.

Nolfi, S., and Floreano, D. (2004). *Evolutionary Robotics*. The MIT Press.

Nolfi, S., and Parisi, D. (1996). Learning to adapt to changing environments in evolving neural networks. *Adapt. Behav.* 5, 75–98. doi: 10.1177/105971239600500104

Nolfi, S., and Tani, J. (1999). Extracting regularities in space and time through a cascade of prediction networks: the case of a mobile robot navigating in a structured environment. *Conn. Sci.* 11, 125–148. doi: 10.1080/095400999116313

Oliva, A. (2005). "Gist of the scene," in *Neurobiology of Attention* (Elsevier), 251–256.

Olivers, C. N., and Roelfsema, P. R. (2020). Attention for action in visual working memory. *Cortex* 131, 179–194. doi: 10.1016/j.cortex.2020.07.011

O'Reilly, R. C. (2020). Unraveling the mysteries of motivation. *Trends Cogn. Sci.* 24, 425–434. doi: 10.1016/j.tics.2020.03.001

Oudeyer, P.-Y., and Kaplan, F. (2008). "How can we define intrinsic motivation," in *Proceedings of the 8th Confenece on Epigenetic Robotics, vol. 5*, 29–31.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evolut. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271

Pace-Schott, E. F., Amole, M. C., Aue, T., Balconi, M., Bylsma, L. M., Critchley, H., et al. (2019). Physiological feelings. *Neurosci. Biobehav. Rev.* 103, 267–304. doi: 10.1016/j.neubiorev.2019.05.002

Palmer, C. R., Barnett, M. N., Copado, S., Gardezy, F., and William, B., Kristan, J. (2014). Multiplexed modulation of behavioral choice. *J. Exp. Biol.* 217(Pt 16), 2963–2973. doi: 10.1242/jeb.098749

Parisi, D., and Petrosino, G. (2010). Robots that have emotions. *Adapt. Behav.* 18, 453–469. doi: 10.1177/1059712310388528

Paus, F., Huang, T., and Asfour, T. (2020). "Predicting pushing action effects on spatial object relations by learning internal prediction models," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (Paris: IEEE), 10584–10590.

Peelen, M. V., and Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *J. Neurosci.* 32, 15728–15736. doi: 10.1523/JNEUROSCI.1953-12.2012

Perry, B., Hogan, L., and Marlin, S. (2000). Curiosity, pleasure and play: a neurodevelopmental perspective. *Haaeyc Advocate* 20, 9–12.

Petras, K., ten Oever, S., Jacobs, C., and Goffaux, V. (2019). Coarse-to-fine information integration in human vision. *Neuroimage* 186, 103–112. doi: 10.1016/j.neuroimage.2018.10.086

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., and Spivey, M. J. (2011). The mechanics of embodiment: a dialog on embodiment and computational modeling. *Front. Psychol.* 2, 5. doi: 10.3389/fpsyg.2011.00005

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., and Spivey, M. J. (2013). Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Front. Psychol.* 3, 612. doi: 10.3389/fpsyg.2012.00612

Pfeifer, R. (2004). "Robots as cognitive tools," in *Cognition and Technology* (John Benjamins Publishing Company), 109–126.

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New view of Intelligence*. MIT Press.

Pfeifer, R., and Scheier, C. (2001). *Understanding Intelligence*. The MIT Press.

Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annu. Rev. Psychol.* 57, 27–53. doi: 10.1146/annurev.psych.56.091103.070234

Pugach, G., Pitti, A., Tolochko, O., and Gaussier, P. (2019). Brain-inspired coding of robot body schema through visuo-motor integration of touched events. *Front. Neurorobot.* 13, 5. doi: 10.3389/fnbot.2019.00005

Puigbo, J.-Y., Pumarola, A., Angulo, C., and Tellez, R. (2015). Using a cognitive architecture for general purpose service robot control. *Conn. Sci.* 27, 105–117. doi: 10.1080/09540091.2014.968093

QueiSSer, J. F., Jung, M., Matsumoto, T., and Tani, J. (2021). Emergence of content-agnostic information processing by a robot using active inference, visual attention, working memory, and planning. *Neural Comput.* 33, 2353–2407. doi: 10.1162/neco_a_01412

Ragozzino, M. E. (2007). The contribution of the medial prefrontal cortex, orbitofrontal cortex, and dorsomedial striatum to behavioral flexibility. *Ann. N. Y. Acad. Sci.* 1121, 355–375. doi: 10.1196/annals.1401.013

Ramirez-Pedraza, R., and Ramos, F. (2021). Decision-making bioinspired model for target definition and "satisfactor" selection for physiological needs. *Cogn. Syst. Res.* 66, 122–133. doi: 10.1016/j.cogsys.2020.10.020

Reed, C. L., and Hartley, A. A. (2021). "Embodied attention: Integrating the body and senses to act in the world," in *Handbook of Embodied Psychology* (Springer), 265–290.

Reiss, S. (2004). Multifaceted nature of intrinsic motivation: the theory of 16 basic desires. *Rev. Gen. Psychol.* 8, 179–193. doi: 10.1037/1089-2680.8.3.179

Rietveld, E. (2008). Situated normativity: the normative aspect of embodied cognition in unreflective action. *Mind* 117, 973–1001. doi: 10.1093/mind/fzn050

Rietveld, E., Denys, D., and Van Westen, M. (2018). "Ecological-enactive cognition as engaging with a field of relevant affordances," in *The Oxford Handbook of 4E Cognition* (Oxford: Oxford University Press), 40–70.

Rolf, M., Steil, J. J., and Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *IEEE Trans. Auton. Ment. Dev.* 2, 216–229. doi: 10.1109/TAMD.2010.2062511

Rolls, E. T. (2000). Précis of the brain and emotion. *Behav. Brain. Sci.* 23, 177–191. doi: 10.1017/S0140525X00002429

Rosenbaum, D. A., Herbort, O., Van der Wel, R., and Weiss, D. J. (2014). What's in a grasp. *Am. Sci.* 102, 366–373. doi: 10.1511/2014.110.366

Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020

Şahin, E., Cakmak, M., Doğar, M. R., Uğur, E., and Üçoluk, G. (2007). To afford or not to afford: a new formalization of affordances toward affordance-based robot control. *Adapt. Behav.* 15, 447–472. doi: 10.1177/1059712307084689

Sainburg, R. L., Lateiner, J. E., Latash, M. L., and Bagesteiro, L. B. (2003). Effects of altering initial position on movement direction and extent. *J. Neurophysiol.* 89, 401–415. doi: 10.1152/jn.00243.2002

Sansone, C., and Harackiewicz, J. M. (2000). *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*. Elsevier.

Sarlegna, F. R., and Sainburg, R. L. (2009). "The roles of vision and proprioception in the planning of reaching movements," in *Advances in Experimental Medicine and Biology* (Springer US), 317–335.

Saunders, J. A. (2004). Visual feedback control of hand movements. *J. Neurosci.* 24, 3223–3234. doi: 10.1523/JNEUROSCI.4319-03.2004

Saunders, J. A., and Knill, D. C. (2003). Humans use continuous visual feedback from the hand to control fast reaching movements. *Exp. Brain Res.* 152, 341–352. doi: 10.1007/s00221-003-1525-2

Schillaci, G., Ciria, A., and Lara, B. (2020a). "Tracking emotions: intrinsic motivation grounded on multi-level prediction error dynamics," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Valparaiso: IEEE), 1–8.

Schillaci, G., Hafner, V. V., and Lara, B. (2016). Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents. *Front. Rob. AI* 3, 39. doi: 10.3389/frobt.2016.00039

Schillaci, G., Villalpando, A. P., Hafner, V. V., Hanappe, P., Colliaux, D., and Wintz, T. (2020b). Intrinsic motivation and episodic memories for robot exploration of high-dimensional sensory spaces. *Adapt. Behav.* 29, 549–566. doi: 10.1177/1059712320922916

Schmidhuber, J. (1991). "A possibility for implementing curiosity and boredom in model-building neural controllers." in *From Animals to Animats* (The MIT Press), 222–227.

Schulkin, J., and Sterling, P. (2019). Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends Neurosci.* 42, 740–752. doi: 10.1016/j.tins.2019.07.010

Schulz, A. (2015). Interoception and stress. *Front. Psychol.* 6, 993. doi: 10.3389/fpsyg.2015.00993

Sereno, M. I., and Huang, R.-S. (2006). A human parietal face area contains aligned head-centered visual and tactile maps. *Nat. Neurosci.* 9, 1337–1343. doi: 10.1038/nn1777

Serino, A. (2019). Peripersonal space (PPS) as a multisensory interface between the individual and the environment, defining the space of the self. *Neurosci. Biobehav. Rev.* 99, 138–159. doi: 10.1016/j.neubiorev.2019.01.016

Sherrington, C. S. (1907). On the proprio-ceptive system, especially in its reflex aspect. *Brain* 29, 467–482. doi: 10.1093/brain/29.4.467

Singh, S., Lewis, R. L., Barto, A. G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Auton. Ment. Dev.* 2, 70–82. doi: 10.1109/TAMD.2010.2051031

Sisbot, E., Alami, R., Simeon, T., Dautenhahn, K., Walters, M., and Woods, S. (2005). "Navigation in the presence of humans," in *5th IEEE-RAS International Conference on Humanoid Robots, 2005* (Tsukuba: IEEE), 181–188.

Smith, C., and Lazarus, R. (1990). *Emotion and Adaptation, Vol. 21, Chapter 17*. Guilford.

Soudry, Y., Lemogne, C., Malinvaud, D., Consoli, S.-M., and Bonfils, P. (2011). Olfactory system and emotion: common substrates. *Eur. Ann. Otorhinolaryngol. Head Neck Dis.* 128, 18–23. doi: 10.1016/j.anorl.2010.09.007

Stagnitti, K. (2004). Understanding play: the implications for play assessment. *Aust. Occup. Ther. J.* 51, 3–12. doi: 10.1046/j.1440-1630.2003.00387.x

Stahl, A. E., and Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science* 348, 91–94. doi: 10.1126/science.aaa3799

Stradner, J., Hamann, H., Schmickl, T., and Crailsheim, K. (2009). "Analysis and implementation of an artificial homeostatic hormone system: a first case study in robotic hardware," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (St. Louis, MO: IEEE), 595–600.

Strigo, I. A., and Craig, A. D. B. (2016). Interoception, homeostatic emotions and sympathovagal balance. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20160010. doi: 10.1098/rstb.2016.0010

Sutton, R., and Barto, A. (1998). Reinforcement learning: an introduction. *IEEE Trans. Neural Netw.* 9, 1054–1054. doi: 10.1109/TNN.1998.712192

Tanaka, H., Ishikawa, T., Lee, J., and Kakei, S. (2020). The cerebro-cerebellum as a locus of forward model: a review. *Front. Syst. Neurosci.* 14, 19. doi: 10.3389/fnsys.2020.00019

Tani, J. (1996). Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Trans. Syst. Man Cybern. B* 26, 421–436. doi: 10.1109/3477.499793

Taormina, R. J., and Gao, J. H. (2013). Maslow and the motivation hierarchy: measuring satisfaction of the needs. *Am. J. Psychol.* 126, 155. doi: 10.5406/amerjpsyc.126.2.0155

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Turner, R. M. (1998). Context-mediated behavior for intelligent agents. *Int. J. Hum. Comput. Stud.* 48, 307–330. doi: 10.1006/ijhc.1997.0173

Tuthill, J. C., and Azim, E. (2018). Proprioception. *Curr. Biol.* 28, R194-R203. doi: 10.1016/j.cub.2018.01.064

Tyng, C. M., Amin, H. U., Saad, M. N. M., and Malik, A. S. (2017). The influences of emotion on learning and memory. *Front. Psychol.* 8, 1454. doi: 10.3389/fpsyg.2017.01454

Van de Cruys, S. (2017). *Affective Value in the Predictive Mind*. Frankfurt am Main: MIND Group.

van Polanen, V., and Davare, M. (2015). Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia* 79, 186–191. doi: 10.1016/j.neuropsychologia.2015.07.010

Vargas, P., Moioli, R., de Castro, L. N., Timmis, J., Neal, M., and Zuben, F. J. V. (2005). "Artificial homeostatic system: a novel approach," in *Advances in Artificial Life* (Berlin; Heidelberg: Springer Berlin Heidelberg), 754–764.

Vargas, P. A., Moioli, R. C., von Zuben, F. J., and Husbands, P. (2009). Homeostasis and evolution together dealing with novelties and managing disruptions. *Int. J. Intell. Comput. Cybern.* 2, 435–454. doi: 10.1108/17563780910982680

Vouloutsi, V., Lallée, S., and Verschure, P. F. M. J. (2013). "Modulating behaviors using allostatic control," in *Biomimetic and Biohybrid Systems* (Berlin; Heidelberg: Springer Berlin Heidelberg), 287–298.

Wade, N. J. (2019). Microscopic anatomy of sensory receptors. *J. Hist. Neurosci.* 28, 285–306. doi: 10.1080/0964704X.2018.1554298

White, R. W. (1959). "Motivation reconsidered: the concept of competence," in *Perspectives in Psychology* (Scott, KS; Foresman & Co), 33–59.

Withagen, R., de Poel, H. J., Araújo, D., and Pepping, G.-J. (2012). Affordances can invite behavior: reconsidering the relationship between affordances and agency. *New Ideas Psychol.* 30, 250–258. doi: 10.1016/j.newideapsych.2011.12.003

Wolpert, D., Ghahramani, Z., and Jordan, M. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882. doi: 10.1126/science.7569931

Wolpert, D., and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Netw.* 11, 1317–1329. doi: 10.1016/S0893-6080(98)00066-5

Wolpert, D. M., Ghahramani, Z., and Flanagan, J. (2001). Perspectives and problems in motor learning. *Trends Cogn. Sci.* 5, 487–494. doi: 10.1016/S1364-6613(00)01773-3

Yoshida, N. (2017). Homeostatic agent for general environment. *J. Artif. Gen. Intell.* 8, 1–22. doi: 10.1515/jagi-2017-0001

Yukie, N. (2011). From affordances to situated affordances in robotics - why context is important. *Front. Comput. Neurosci.* 5, 30. doi: 10.3389/conf.fncom.2011.52.00030

Zambelli, M., Cully, A., and Demiris, Y. (2020). Multimodal representation models for prediction and control from partial information. *Rob. Auton. Syst.* 123, 103312. doi: 10.1016/j.robot.2019.103312

Zenha, R., Vicente, P., Jamone, L., and Bernardino, A. (2018). "Incremental adaptation of a robot body schema based on touch events," in *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Tokyo: IEEE), 119–124.

Zhang, Q., and Li, S. (2019). The roles of spatial frequency in category-level visual search of real-world scenes. *PsyCh J.* 9, 44–55. doi: 10.1002/pchj.294

Zhang, T., Hu, F., Deng, Y., Nie, M., Liu, T., Wu, X., et al. (2018). "Self-developing proprioception-based robot internal models," in *Intelligence Science II* (Springer International Publishing), 321–332.

Ziemke, T., and Lowe, R. (2009). On the role of emotion in embodied cognitive architectures: from organisms to robots. *Cognit. Comput.* 1, 104–117. doi: 10.1007/s12559-009-9012-0

Zimmermann, M., Meulenbroek, R. G., and de Lange, F. P. (2012). Motor planning is facilitated by adopting an action's goal posture: an fmri study. *Cereb. Cortex* 22, 122–131. doi: 10.1093/cercor/bhr098

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Valenzo, Ciria, Schillaci and Lara. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

PERSPECTIVE
published: 21 June 2022
doi: 10.3389/fnbot.2022.897110

# Am I (Deep) Blue? Music-Making AI and Emotional Awareness

*Nicholas Novelli*[1]* *and Shannon Proksch*[2]

[1] Independent Researcher, Winnipeg, MB, Canada, [2] Cognitive and Information Sciences, University of California, Merced, Merced, CA, United States

Artificial Intelligence has shown paradigmatic success in defeating world champions in strategy games. However, the same programming tactics are not a reasonable approach to creative and ostensibly emotional artistic endeavors such as music composition. Here we review key examples of current creative music generating AIs, noting both their progress and limitations. We propose that these limitations are rooted in current AIs lack of thoroughly embodied, interoceptive processes associated with the emotional component of music perception and production. We examine some current music-generating machines that appear to be minimally addressing this issue by appealing to something akin to interoceptive processes. To conclude, we argue that a successful music-making AI requires both the generative capacities at which current AIs are constantly progressing, and thoroughly embodied, interoceptive processes which more closely resemble the processes underlying human emotions.

**Keywords: music cognition, artificial intelligence, emotion, interoception, creativity, aesthetics**

OPEN ACCESS

***Edited by:***
*Adam Safron,*
*Johns Hopkins Medicine,*
*United States*

***Reviewed by:***
*Karl Friston,*
*University College London,*
*United Kingdom*

***Correspondence:**
*Nicholas Novelli*
*nicholas.a.novelli@gmail.com*

***Received:*** *15 March 2022*
***Accepted:*** *20 May 2022*
***Published:*** *21 June 2022*

***Citation:***
*Novelli N and Proksch S (2022) Am I (Deep) Blue? Music-Making AI and Emotional Awareness. Front. Neurorobot. 16:897110. doi: 10.3389/fnbot.2022.897110*

## 1. INTRODUCTION

In the race to build increasingly autonomous–perhaps even conscious–machines, focus on machine learning and machine intelligence is on the rise. Paradigmatic AI successes in games such as chess and Go have relied heavily on computational processes that occur primarily "in the head" of game-playing agents. 4E (embodied, embedded, enactive, extended) approaches to cognition are increasingly demonstrating the importance of cognitive processes which extend beyond such rule-based symbol manipulation, and into the bodies and external environments of cognitive agents. The next great frontier for autonomous intelligent systems is human creativity and art. Specifically, an art form that encapsulates the tenets of 4E cognition and places an emphasis on the agent's interaction with their social environment, as well as their external and internal milieu: music. Numerous music-making AIs have been created in attempts to simulate, understand, or replicate the process of human creativity in musical composition using artificial neural networks, such as Google's Magenta, Cambridge University's BachBot, or Sony CSL's Flow Composer.

There are practical reasons that computers have difficulty performing creative tasks as successfully as strategic tasks, due to both mathematical complexity and a deep connection with emotional processing in human music-making. These emotional processes have roots in bodily, physiological, and autonomic states in the performer and the listener. We draw on theories that emphasize thoroughly embodied, interoceptive processes rooted in the prediction and regulation of internal physiological processes as part of the mechanism of human emotion (Seth and Friston, 2016), and extend these theories to musical perception and production (Proksch, 2018). If these bodily processes are crucial to creative musical success, then AIs will need such mechanisms

(or analogues to them) in order to create authentic music. We argue that generative music AIs must experience, or robustly simulate, something akin to the interoceptive processes that underlie emotional states.

# 2. ARTIFICIAL INTELLIGENCE: DEFEATING CHAMPIONS, APPROXIMATING MUSIC

## 2.1. AI Successes: Defeating Champions

### 2.1.1. DeepBlue and AlphaGo

If AI Success is measured in terms of the ability to equal (or outperform) expert human counterparts, then IBM's Deep Blue and Google's AlphaGo are paradigmatic successes. DeepBlue beat Gary Kasparov in 1997 by mapping every possible combination of moves it could make, up to the next six moves, according to a set of pre-programmed rules and evaluations established with the help of expert chess players (Campbell, 1998). Nearly 20 years later, AlphaGo beat Lee Sedol, the world champion of the strategy game "Go."

DeepBlue and AlphaGo both used a directed graph, called a game tree, which represents possible moves and positions for multiple sequences of game play. A complete game-tree for chess would contain $10^{120}$ moves (already more than the number of atoms in the universe, $10^{80}$), and a game-tree for Go would massively exceed this number. Before AlphaGo, most Go AIs achieved this computational feat by using a technique called Monte Carlo Tree Search. Instead of looking through pre-programmed mappings of every possible combination of moves from a current state, the AI stores only the rules, and runs multiple simulations extending from the current state until any winning state, re-running this simulation for every move. AlphaGo used machine learning combined with policy and value (neural) networks to determine the move with the highest likelihood of a win, in combination with simulative tree-search methods (Silver et al., 2016). Using a variety of methods, strategy game AIs have achieved expert-level success. Given AI successes at strategy games, one might wonder if these techniques extend to creative (albeit rule-based) endeavors such as music-making. Could this strategy work for music? Music composition has far more possibilities at each decision-point than chess or Go, and does not have rules in the same sense of options being forbidden and thus eliminated (many of our most beloved songs break a supposed "rule" of composition; Cochrane, 2000), nor success or failure states that end the exercise early.

It is impractical to even store a "tree" of all possible musical sequences of any reasonable length, let alone to encode any information about their aesthetic viability. A song might range across two or more octaves, each with twelve notes and 60 possible chords for each instrument part, plus differences in rhythm, tempo, orchestration, to say nothing of going beyond the traditional Western paradigm, e.g., including quarter-tones. We need a shortcut. Current music-making AIs have attempted to achieve this shortcut in ways similar to AlphaGo, using recurrent neural networks with value programmed as the probabilistic likelihood of the next note, given the structural information of the previous notes in terms of rhythm, pitch, etc. These AIs have produced promising results (as we shall see) but the evidence is growing that there is a limit to what can be accomplished with probabilistic structural data of the musical input alone.

## 2.2. AI Works in Progress: Approximating Music

Music-generating AIs work primarily by learning patterns in the structural information-pitch, rhythm, harmony, etc. of a set of musical training data. After training, the AI is provided with a set of rules and some form of starting cue from which it generates a piece of music. There are many different methods one can use to create a music generating AI, reviewed in detail in Carnovalini and Rodà (2020). Here, we will focus on three examples of generative AIs which learn from musical training input to probabilistically generate musical compositions.

### 2.2.1. Magenta

Magenta is a far-reaching Google project that seeks to determine whether machine learning can be used to create "compelling art and music"[1]. Magenta's early music compositions used Recurrent Neural Networks (RNNs), which work by learning a probability distribution of possible inputs given previous data, in order to predict the next input. After a series of training data, the RNN can now generate its own output using the same probabilistic rules. Magenta is trained on thousands of monophonic (single note at a time) melodies, from which it learns the rules and style of those melodies, and develops probabilistic models which it uses to generate new monophonic melodies on its own.

The result, in Magenta's first ever composition in June 2016, was a fairly impressive, albeit simplistic composition, reminiscent of a standard theme and variations[2]. Just under one and one-half minutes long, it begins with a simple but clear motif set to a repeating eighth-note and quarter-note rhythm, which repeats first verbatim, then again with a bit of clumsy ornamentation, before entering a "creative" development section introducing some awkward new rhythms. The piece returns to the original motif, before ultimately and abruptly stopping mid-phrase with no clear conclusion.

This abrupt ending occurs because, unlike in a human composition, many music-making AIs thus far have no concept of a musical "narrative arc" with conclusive resolutions. Instead, they will tend to "wander around" their music-making process unless a human programs it to stop at a certain point. Much of Magenta's work thus far has dealt with monophonic compositions, generating a single note at a time[3]. This contrasts with the next examples that use multiple streams of notes.

[1]Magenta is an open source project from the Google Brain team.Source code and updates can be found at: https://github.com/tensorflow/magenta.

[2]To listen to Magenta's first composition: https://www.youtube.com/watch?v=lht-emTioLw.

[3]There has been improvement in establishing a narrative structure by implementing hierarchical models such as MusicVAE. MusicVAE implements a variational autoencoder consisting of a bidirectional LSTM network and a novel hierarchical RNN decoder. This allows for generation of music sequences with a more coherent long-term structure. Listen to MusicVAE at: https://magenta.tensorflow.org/music-vae.

#### 2.2.2. BachBot

At first listen, BachBot's[4] compositions are much more impressive than the debut single-note melody from the Magenta Project. BachBot also uses Long-Short Term Memory (LSTM) RNNs, and is trained specifically on Bach chorales. Unlike Magenta, its training input is homophonic, or chordal, with a series of simultaneous pitches (chords) organized as melody plus harmony, all composed by J.S. Bach. Like Magenta, BachBot used its probabilistic models of which chords should come next to generate new pieces (here in the style of Bach chorales) on its own. The music created by this AI is much more sophisticated than the melodic-play generated by Magenta, and to the untrained ear is virtually indistinguishable from that composed by Bach himself[5].

Unlike Magenta's rough approximation of a theme and variations, these chorales follow a cadential structure with well-organized phrases[6]. Instead of awkwardly-placed ornaments, BachBot's chorales contain purposeful passing tones within a stable rhythmic structure. However, while impressive, BachBot suffers some of the same problems as Magenta. Unless Bachbot is given at least one line of a chorale (or a melody) to harmonize over, it will suffer the same "wandering" fate as Magenta's compositions. It maintains its semblance of structure because a human provides it with a prescribed line of notes, which then constrains its output and leads BachBot through a structured journey of composition.

#### 2.2.3. Flow Composer (Paris, Sony Computer Science Laboratories, ERC Funded Project

The even more impressive Flow Composer[7] was created to produce pop songs. Rather than LSTM's, Flow Composer uses Markov constraints. This solves the "wandering" problem faced by both Magenta and Bachbot by generating finite-length sequences, and similar to Bachbot it generates these sequences in accordance with a given composer-style, or genre of music. Flow Composer takes input for model generation in the form of lead sheets (basic chord structure plus a melody line), and once again uses its probabilistic memory to generate a new lead sheet for a new song all on its own.

Flow Composer created the first ever full-length pop song composed by an AI, the Beatles-"inspired" track "Daddy's Car"[8]. Daddy's Car has lyrics, with multiple voices, guitar, drums—a full orchestration. However, as impressive as this is, and despite the problems apparently solved from Magenta and Bachbot dealing with wandering and improper ornamentation or rhythm, the only thing that Flow Composer generates is a lead sheet. The rest of the music composing process, including writing the harmonies themselves, instrumentation, and writing the lyrics, are performed by human collaborators[9].

### 2.3. What's Missing?

There is a common thread amongst these music-making AIs, and that is the importance of the human in the process. This is partially rooted in the fact that each example is not truly generating musical content, but is reliant to some degree on human intervention. In fact, Magenta, with the most basic and least impressive of the compositions highlighted here, composes music with the least help from human musical decisions. If each composition was subjected to a sort of musical "Turing test," Magenta's might be the least likely to pass because it rests in an "uncanny valley" between quality music and childlike—or just plain strange—artificial creativity. However, the other compositions might pass solely due to the human intervention necessary to yield the final musical product. It might be countered that this is simply a difference of degree, because human composers are still better at music than our AIs. But at what, exactly, are humans better?

Although music students and young musicians are taught and trained in the rules and norms of their musical culture, there is a common pre-theoretical or folk-psychological notion that what is important in composing music is the expression or elicitation of emotion. Good music does not just blindly follow rules, it has feeling, emotion. Historical and current work in music cognition indicates that part of what enables humans to both process and create music in the way that we do involves inherently emotional processes (Huron, 2008; Juslin and Västfjäll, 2008; Trost et al., 2012; Koelsch et al., 2015). Current trends in the philosophy of cognitive science indicate these emotional processes are rooted in the prediction and regulation of internal physiological processes, or interoceptive states. Conceiving of the experience of emotional states as crucially involving interoceptive processing has important implications for music-making AIs.

## 3. EMOTION AND INTEROCEPTION

There are competing accounts of what makes an "emotion," however all accounts consider the importance or interoceptive, physiological states of the body. If an emotional experience arises from gathering evidence from the state of our body, plus a subsequent-or simultaneous-cognitive appraisal (James, 1884; Lange, 1885; Schachter and Singer, 1962), then the brute-force rule-following and simulation-based success of strategy game AIs could be extrapolated to emotional and creative processes like music making. There are some reliable cross cultural mappings of particular musical sounds to particular (potentially emotional) functions—such as the downward melodic passages and slow

[4]BachBot is also open source, run by researchers at the University of Cambridge and Microsoft Research Center. You can find source code and updates at: https://github.com/feynmanliang/bachbot.

[5]You can listen to BachBot at: https://soundcloud.com/bachbot.

[6]Performance RNN generates polyphonic compositions of solo piano, and incorporates expressive timing by encoding a flexible rather than strict metrical grid while also allowing the dynamics of each note to vary. Listen to Performance RNN at: https://magenta.tensorflow.org/performance-rnn.

[7]FlowMachines, by Sony Computer Science Laboratories in Paris, is a European Research Council funded project. Unfortunately, it is not an open source project. However, more information on their project, including DeepBach (FlowMachines take on BachBot) can be found at: http://www.flow-machines.com/.

[8]You can listen to "Daddy's Car" at: https://www.youtube.com/watch?v=LSHZ_b05W7o.

[9]Remember, the melody is incorporated as part of the generated leadsheet, as are the instructions for which notes to include in the harmonies, but not necessarily which voice those notes should be assigned to the harmonies—e.g., the guitar or the bass.

rhythms of soothing lullabies (Mehr et al., 2019). Indeed, musicians make use of standard motifs within their musical traditions that are associated with or meant to evoke certain emotions in an audience. An AI could, in theory, form a reliable mapping between statistical regularities of music and emotion across cultures, even without a physical body to instantiate those interoceptive processes itself.

However, making music that elicits or evokes an appropriate emotion is not as simple as choosing from a library of sound sequences coded for emotional content. Emotional experience relies on expectations about the way that interoceptive states of the body will unfold with respect to the external and social context of that experience (Critchley, 2005; Seth, 2013; Seth and Critchley, 2013). In this vein, making music with emotion relies on expectations about the way that interoceptive states of the body will respond to music-listening and music-making. In fact, experience with the bodily movement involved in making and moving music leads to enhanced interoceptive awareness for both musicians and dancers alike (Schirmer-Mokwa et al., 2015; Christensen et al., 2018). Information from relevant interoceptive states (whether first-hand, i.e., having a body capable of them, or second-hand, i.e., interacting with an individual that does), can enhance artificial music generation systems' ability to create compelling music.

The experience of emotion in music listening and music production is rooted in expectancy of not just the structural information of music (for which our AIs are very capable), but also more thoroughly embodied expectancy of the internal, physiological state which is either cued or expressed in music listening or creating, respectively (Proksch, 2018). Music listening is a common tool by which individuals monitor and regulate their emotional states and supporting neurochemistry (Chanda and Levitin, 2013). Consider the practice of listening to calming music as you fall asleep—calming because it cues the brain to minimize levels of cortisol and adrenaline in your body (McKinney et al., 1997; Khalfa et al., 2003; Thoma et al., 2013). Or the opposite, listening to upbeat energetic music on your morning run, inciting increased general arousal, and enjoyment of physical exertion (jymming', c.f. Fritz et al., 2013a,b), while the exercise itself may even boost increased enjoyment of the music (Hove et al., 2021). These same processes are leveraged by a composer, or an improviser, who is creating music in response to or in order to modulate their audience's emotional, and by extension physiological, states.

The exteroceptive information of the music, the structural organization of pitch and rhythm, is mutually contextualized by the thoroughly embodied, interoceptive information which is used to generate the music itself, and the integration of these two forms of information in the creative process of music composition is what leads to the pre-theoretical intuition that may be deemed the "heart and soul" of a musical work. Music-making AIs, and the music they compose, are thought to lack this emotional quality. Since computer programs lack the proper embodied, interoceptive states and homeostatic physiological drive, by which emotions are proposed to be constituted, then this pre-theoretical intuition is plainly justified. It seems that music-making AIs cannot compose authentic music because they lack the interoceptive, emotional processes necessary to do so.

Thus far, we have observed that the most musically impressive programs have a higher degree of human intervention to achieve a satisfying musical structure. We have justified criticisms of AI's musical output by demonstrating that the pre-theoretical notion that computer composed music lacks the "heart and soul" of authentic, human composed music can be rooted in basic interoceptive processes of physiological homeostasis. By incorporating this more thoroughly embodied process, AIs may be able to avoid their worst tendencies and ground their musical output in terms of their own interoceptive states, and come closer to attaining states resembling something like human emotions. In this next section, we will present two more music machines which may create closer approximations to authentic musical works.

## 3.1. Minimally Interoceptive Artificial Music Generation

### 3.1.1. Magenta: AI Duet

We return to the Magenta project to visit an interactive, improvising music machine. AI Duet runs on similar models as we've discussed earlier, using RNNs and LSTMs to learn the rules and styles of its input, and then using those rules to generate its own musical output. However, this time, AI Duet takes input from a human, improvising musician—in real time—and together they improvise their own joint musical performance. Simply put, when you play a series of notes, the computer will respond to those notes, sometimes mimicking, mirroring, or expanding on the input you've given it. This is much more impressive than Magenta's initial attempts at music composition, in that it incorporates real time social interaction. The music generated by these social performances range from an awkward situation between two mediocre or completely inexperienced improvisers, to fairly convincing collaborative experiments. These experiments, whether or not they are aesthetically pretty or pleasing musically, have some semblance of feeling. This is because the musical event as a whole, in this case, is partially rooted in interoceptive processes—albeit only the embodied interoceptive processes of the human collaborator. The computer program itself is still only processing the exteroceptive content of the musical structure. If two of these music AIs interact together, the music produced quickly becomes nonsensical in the same way that the conversation between two chatbots quickly deteriorates.

### 3.1.2. Cybraphon

Designed in 2009 by the FOUND artist collective together with Simon Kirby from the University of Edinburgh, Cybraphon is a "moody, autonomous robot band in a box" and is housed at the National Museum of Scotland (National Museums Scotland, 2009). The instrument is quite literally a wardrobe, filled with musical objects, lights, an "emotion meter," and a computer which controls when each of these objects will sound, light up, or move (Taubman, 2014). Unlike the previous AIs discussed, Cybraphon is not entirely generative from the ground up. Rather, it performs by choosing from a repertoire of precomposed bits of music, and selects the music that corresponds to its current "emotional" state. This "emotional state" is not a product of

solving the problem of giving a robot a homeostatic body, but rather from being thoroughly entrenched in social media. Cybraphon is a bit of a diva—it Googles itself every fifteen seconds and observes its current popularity over news sites, twitter, and facebook. The more online activity, and the more positive the online activity, Cybraphon will "cheer up" and might play one of its happy tunes. If no online interaction is happening, it will sit in a state of perpetual indifference, refusing to make music at all[10]. Cybraphon has something like extended interoceptive processing, loosely embodied by activity of the online community. Although similar to AI Duet's reliance on human interoceptive processes in the creation of a musical event, Cybraphon does not rely on any one person or group of persons' interoceptive processes, but rather translates social media activity into loosely embodied "emotional" states based on the online activity's deviation from normal levels. This nearly resembles the sort of embodied process in which interoceptive emotions are proposed to be rooted. However, the instrument lacks a predictive component that might enable it to probabilistically seek a homeostatic set point for these extended interoceptive states.

### 3.2. What's Missing

Crucially, the prediction and regulation of interoceptive states relies to some extent on bodily action. In fact, the very language of *seeking* a homeostatic set point to bring about an interoceptive and emotional state implies that an embodied music-making AI must be able to take action in the world to affect its own internal states. Magenta AI:Duet, while minimally interactive, does not have a body to take action in the world or interoceptive processes to respond to the rhythmic and melodic content that is co-generated by the AI and human performer. It relies on the actions of the human duet partner. Cybraphon does take some action in the world through the small repertoire of mechanical actions it can make in response to its extended interoceptive state, which itself depends on engagement of others in the world. However, Cybraphon's own actions have no effect on its interoceptive states, and it cannot interact with other individuals during its music making. The ability to act on interoceptive processes, and interact with other individuals, may be one more ingredient missing for successful music generation by music-making machines.

## 4. ACTION AND INTERACTION

### 4.1. Movement in Music Generating Robots

Human music-making is, itself, "inseparable" from movement (Keller and Rieger, 2009). Even passive music listening is strongly rooted in motor processes in the brain (Grahn and Brett, 2007; Gordon et al., 2018). Anticipation of melodic, harmonic, and rhythmic content of a musical work engages canonical emotion, reward, and motor networks in the brain (Salimpoor et al., 2015; Vuust et al., 2022). Rhythmic components of music are acutely associated with predictive and motor processes (Koelsch et al., 2019; Proksch et al., 2020). In particular, there is a human urge to move to a musical beat that may be strongly related to the balance of sensory prediction and prediction error elicited by rhythmic syncopation found in musical groove, and higher levels of musical groove are rated as more pleasurable (Janata et al., 2012; Witek et al., 2014). Joint movement to musical rhythm can result in the co-activation of motor networks related to the perception of self and other (Overy and Molnar-Szakacs, 2009; Friston and Frith, 2015), engaging the endogenous opioid system and mirroring mechanisms which support social bonding (Tarr et al., 2014). There has been increasing recognition of music as an inherently enactive and interactive process, mutually co-constituted in the actions of musicians predicting both musical (exteroceptive) and bodily sensations (interoceptive and proprioceptive states) (Cross, 2014; Dell'Anna et al., 2021). Joint musical interaction is further aided by visual (exteroceptive) information regarding the movement, intention, and interest of each musician in addition to internal representations of movements of the other interacting musicians (Novembre et al., 2012, 2014). We next provide an example of an embodied and interactive music-making AI—an improvising, marimba-playing robot.

### 4.2. Shimon

Created by the Robotic Musicianship Group at Georgeia Tech Center for Musical Technology, Shimon is trained on an extensive repertoire of classical, jazz, and popular music. Similar to Magenta AI:Duet, Shimon is a music-making AI that engages in musical improvisation alongside human performers. However, Shimon is physically embodied in a marimba playing robot with four arms that can play melodic, rhythmic, and multiphonic music (Weinberg et al., 2009). Further, this robot features an expressive "face" that can move along with a musical beat and facilitate interaction with ensemble musicians[11]. Similar embodied robotic music machines have been designed to play traditional instruments, such as piano, violin, and flute as well as new forms of musical instruments afforded by the different physical configurations a musical robot can take compared to humans (Bretan and Weinberg, 2016). These robots create "a sense of embodiment" that afford "richer musical interactions" between human musicians and robotic music-making AIs (Bretan and Weinberg, 2016). For instance, Shimon nods along to the musical beat—mimicking the human propensity to move to a beat—and direct its attention by turning its head toward the musician playing the most interesting (i.e., salient) musical line when interacting in a musical ensemble (Weinberg et al., 2009). The actions taken by the robots may not refer to any interoceptive or emotional state internal to the robot, but may be robustly simulating such states through their actions so as to facilitate self-other merging and social emotions among their human co-performers.

[10]Listen to Cybraphon at: https://www.youtube.com/watch?v=wDyabLAzKuo.

[11]Learn more about Shimon and listen to some of its music at: https://www.shimonrobot.com/.

# 5. DISCUSSION AND CONCLUSION

## 5.1. Additional Considerations: Musico-Historical and Social Context

There are other aesthetic properties that an even more embodied, embedded, environmentally interactive form of AI might succeed at producing. Music can be judged not only by how it makes us feel and how pleasant it sounds, but on the basis of properties like innovativeness, subversiveness, homage to other works, etc. For an AI to master these properties, it would have to have an awareness of musico-historical context, beyond mere probability distributions over the notes, rhythms, and features of a particular song or musical style. Jerrold Levinson enumerates musico-historical context as some personal components (a composer's own style, repertoire, oeuvre, and influences), and some general components (the history of musical development, prevalent musical styles, and influences at time of composition, and activities of contemporary composers) (Levinson, 1980). A musical AI might need to be socially embedded within a musico-historical context to have mastery of these complex—and even some more simple—aesthetic properties. A composer does not rely on her own feelings alone, and imagine if Cybraphon could not only monitor social media reactions, but also processed the nature of positive and negative critiques and tracked exactly which aspects of its compositions some listeners find annoying or sublime. But a more robust system wherein an A.I. composer is educated by an artistic community could develop the ability to create beyond its teachers or its training data, and play off the works of others in a way that adds aesthetic depth[12].

All this might require engaging and participating in a musical community rather than simply processing data.

### 5.1.1. Spawn

One such system has been created by Holly Herndon. The singing AI called "Spawn" was trained on her own voice, the voices of her musical collaborators (Friedlander, 2019), and even the voices of her audience (Herndon, 2019). Herndon says that as opposed to AIs such as Bachbot that make music in one particular style, her goal was to create an AI that can "understand the logic of a sound sample" and thus be more adaptable (Friedlander, 2019). A strong emphasis of the project is the "raising" of an AI by a "community." Herndon is careful to be transparent about the limitations of the technology and estimates the contribution of the AI in each musical composition at about twenty percent.

## 5.2. The Future of Music Machines

To conclude, while music-making AI is thriving on the progress we have made in generative music machines, something is yet missing. Music-making AIs are unable to reach success by relying on brute-force rule memorization and future state simulation in the same manner as competitive strategy game AIs. Successful compositions by music-making AIs thus far, while appearing autonomously generative, have required a significant amount of human intervention. Even with this intervention, these compositions seem to be lacking feeling, emotion, and a focused narrative structure. We demonstrated that human music production and perception is not merely "in the head," but rather involves influence from homeostatic, interoceptive processes in which human emotion processing is grounded. This interoceptive processing is importantly lacking in computer programs creating musical compositions. Two music-machines, AI Duet and Cybraphon appear to be minimally incorporating a form of interoceptive processing, however the former is reliant upon input from a human collaborator, and the latter is not generative. Current music-making robots, such as Shimon, may be more adept at mimicking actions which, when made by a human, are rooted in emotional and interoceptive processes—enabling rich musical interactions as a member of a musical ensemble. Spawn is an example of a budding musical AI which is raised by and embedded in a community, learning, and evolving through interactions with humans rather than from pre-composed datasets of music. A successful music-making AI will need to build on current generative successes, and incorporate more thoroughly embodied interoceptive processing of a sort that would ground the machine's musical output to its own internal, perhaps even conscious, states. Essentially, they must be able to ask themselves, "Am I Blue?."

# DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

# AUTHOR CONTRIBUTIONS

NN contributed expertise on artificial intelligence, cyborg technology, aesthetics, and philosophy of music. SP contributed expertise on music cognition, 4e processing, interoception, and machine learning and neural networks. Both authors contributed to writing and revision of all aspects of the manuscript, and approved the submitted version.

[12] An important consideration, though beyond the scope of this article to discuss in detail, is that designers have seemingly assumed that machine learning is the correct approach to music-making AIs—indeed, all the AIs discussed here rely primarily or exclusively on machine learning. However, for activities that significantly involve complex mutual prediction (this is true of musical collaboration as well as solo musical improvisation and even to some extent non-improvisational solo performance), "learning" in this sense may not be the correct computational approach and *inference*, relying on state estimation and Bayesian filtering, may be a more appropriate paradigm that has been as-yet under-explored (we thank an anonymous reviewer for bringing up this point).

# REFERENCES

Bretan, M., and Weinberg, G. (2016). A survey of robotic musicianship. *Commun. ACM* 59, 100–109. doi: 10.1145/2818994

Campbell, M. S. (1998). *An Enjoyable Game: How HAL Plays Chess. HAL's Legacy*. Cambridge, MA: MIT Press. 75–98.

Carnovalini, F., and Rodá, A. (2020). Computational creativity and music generation systems: an introduction to the state of the art. *Front. Artif. Intell.* 3:14. doi: 10.3389/frai.2020.00014

Chanda, M. L., and Levitin, D. J. (2013). The neurochemistry of music. *Trends Cogn. Sci.* 17, 179–193. doi: 10.1016/j.tics.2013.02.007

Christensen, J. F., Gaigg, S. B., and Calvo-Merino, B. (2018). I can feel my heartbeat: dancers have increased interoceptive accuracy. *Psychophysiology* 55:e13008. doi: 10.1111/psyp.13008

Cochrane, R. (2000). Playing by the rules: a pragmatic characterization of musical performances. *J. Aesthet. Art Crit.* 58, 135–142. doi: 10.1111/1540-6245.jaac58.2.0135

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *J. Compar. Neurol.* 493, 154–166. doi: 10.1002/cne.20749

Cross, I. (2014). Music and communication in music psychology. *Psychol. Mus.* 42, 809–819. doi: 10.1177/0305735614543968

Dell'Anna, A., Leman, M., and Berti, A. (2021). Musical interaction reveals music as embodied language. *Front. Neurosci.* 15:667838. doi: 10.3389/fnins.2021.667838

Friedlander, E. (2019). *Holly Herndon and Her AI Baby Spawned a New Kind of Folk Music*. The Fader. Available online at: https://www.thefader.com/2019/05/21/holly-herndon-proto-ai-spawn-interview

Friston, K. and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405. doi: 10.1016/j.concog.2014.12.003

Fritz, T. H., Halfpaap, J., Grahl, S., Kirkland, A., and Villringer, A. (2013a). Musical feedback during exercise machine workout enhances mood. *Front. Psychol.* 4:921. doi: 10.3389/fpsyg.2013.00921

Fritz, T. H., Hardikar, S., Demoucron, M., Niessen, M., Demey, M., Giot, O., et al. (2013b). Musical agency reduces perceived exertion during strenuous physical performance. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17784–17789. doi: 10.1073/pnas.1217252110

Gordon, C. L., Cobb, P. R., and Balasubramaniam, R. (2018). Recruitment of the motor system during music listening: an ale meta-analysis of fMRI data. *PLoS ONE* 13:e0207213. doi: 10.1371/journal.pone.0207213

Grahn, J. A., and Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *J. Cogn. Neurosci.* 19, 893–906. doi: 10.1162/jocn.2007.19.5.893

Herndon, H. (2019). *Holly Herndon: Birthing PROTO [video]*. Available online at: https://www.youtube.com/watch?v=v_4UqpUmMkg (accessed March 10, 2022).

Hove, M. J., Martinez, S. A., and Shorrock, S. R. (2021). Physical exercise increases perceived musical pleasure: modulatory roles of arousal, affect, or dopamine? *Psychol. Mus.* doi: 10.31234/osf.io/gy6qd

Huron, D. (2008). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.

James, W. (1884). What is emotion? *Mind.* 9:188–205.

Janata, P., Tomic, S. T., and Haberman, J. M. (2012). Sensorimotor coupling in music and the psychology of the groove. *J. Exp. Psychol.* 141:54. doi: 10.1037/a0024208

Juslin, P. N., and Västfjäll, D. (2008). Emotional responses to music: the need to consider underlying mechanisms. *Behav. Brain Sci.* 31, 559–575. doi: 10.1017/S0140525X08005293

Keller, P. E., and Rieger, M. (2009). Musical movement and synchronization. *Music Percept.* 26, 397–400. doi: 10.1525/mp.2009.26.5.397

Khalfa, S., Dalla Bella, S., Roy, M., Peretz, I., and Lupien, S. J. (2003). Effects of relaxing music on salivary cortisol level after psychological stress. *Ann. N.Y. Acad. Sci.* 999, 374–376. doi: 10.1196/annals.1284.045

Koelsch, S., Jacobs, A. M., Menninghaus, W., Liebal, K., Klann-Delius, G., Von Scheve, C., et al. (2015). The quartet theory of human emotions: an integrative and neurofunctional model. *Phys. Life Rev.* 13, 1–27. doi: 10.1016/j.plrev.2015.03.001

Koelsch, S., Vuust, P., and Friston, K. (2019). Predictive processes and the peculiar case of music. *Trends Cogn. Sci.* 23, 63–77. doi: 10.1016/j.tics.2018.10.006

Lange, C. G. (1885). "The mechanism of the emotions," in *The Classical Psychologists*, 672–684. Available online at: https://www.sapili.org/subir-depois/en/ps000132.pdf

Levinson, J. (1980). What a musical work is. *J. Philos.* 77, 5–28. doi: 10.2307/2025596

McKinney, C. H., Antoni, M. H., Kumar, M., Tims, F. C., and McCabe, P. M. (1997). Effects of guided imagery and music (GIM) therapy on mood and cortisol in healthy adults. *Health Psychol.* 16:390. doi: 10.1037/0278-6133.16.4.390

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., et al. (2019). Universality and diversity in human song. *Science* 366:eaax0868. doi: 10.1126/science.aax0868

National Museums Scotland (2009). *Cybraphon*. National Museums Scotland. Retrieved from: https://www.nms.ac.uk/explore-our-collections/stories/science-and-technology/cybraphon/

Novembre, G., Ticini, L. F., Schütz-Bosbach, S., and Keller, P. E. (2012). Distinguishing self and other in joint action. Evidence from a musical paradigm. *Cereb. Cortex* 22, 2894–2903. doi: 10.1093/cercor/bhr364

Novembre, G., Ticini, L. F., Schütz-Bosbach, S., and Keller, P. E. (2014). Motor simulation and the coordination of self and other in real-time joint action. *Soc. Cogn. Affect. Neurosci.* 9, 1062–1068. doi: 10.1093/scan/nst086

Overy, K., and Molnar-Szakacs, I. (2009). Being together in time: musical experience and the mirror neuron system. *Music Percept.* 26, 489–504. doi: 10.1525/mp.2009.26.5.489

Proksch, S. (2018). Interoceptive inference and emotion in music: integrating the neurofunctional quartet theory of emotion-with predictive processing in music-related emotional experience. *J. Cogn. Neuroethics* 5, 101–125. Available online at: http://jcn.cognethic.org/jcnv5i1_Proksch.pdf

Proksch, S., Comstock, D. C., Médé, B., Pabst, A., and Balasubramaniam, R. (2020). Motor and predictive processes in auditory beat and rhythm perception. *Front. Hum. Neurosci.* 14:578546. doi: 10.3389/fnhum.2020.578546

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., and McIntosh, A. R. (2015). Predictions and the brain: how musical sounds become rewarding. *Trends Cogn. Sci.* 19, 86–91. doi: 10.1016/j.tics.2014.12.001

Schachter, S., and Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* 69:379. doi: 10.1037/h0046234

Schirmer-Mokwa, K. L., Fard, P. R., Zamorano, A. M., Finkel, S., Birbaumer, N., and Kleber, B. A. (2015). Evidence for enhanced interoceptive accuracy in professional musicians. *Front. Behav. Neurosci.* 9:349. doi: 10.3389/fnbeh.2015.00349

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Seth, A. K., and Critchley, H. D. (2013). Extending predictive processing to the body: emotion as interoceptive inference. *Behav. Brain Sci.* 36:227. doi: 10.1017/S0140525X12002270

Seth, A. K. and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. B* 371:20160007. doi: 10.1098/rstb.2016.0007

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Tarr, B., Launay, J., and Dunbar, R. I. (2014). Music and social bonding: "self-other" merging and neurohormonal mechanisms. *Front. Psychol.* 5:1096. doi: 10.3389/fpsyg.2014.01096

Taubman, A. (2014). "Cybraphon: collecting the physical or the digital at national museums Scotland?," in *Electronic Visualisation and the Arts (EVA 2014)* (London), 122–128. doi: 10.14236/ewic/EVA2014.31

Thoma, M. V., La Marca, R., Brönnimann, R., Finkel, L., Ehlert, U., and Nater, U. M. (2013). The effect of music on the human stress response. *PLoS ONE* 8:e70156. doi: 10.1371/journal.pone.0070156

Trost, W., Ethofer, T., Zentner, M., and Vuilleumier, P. (2012). Mapping aesthetic musical emotions in the brain. *Cereb. Cortex* 22, 2769–2783. doi: 10.1093/cercor/bhr353

Vuust, P., Heggli, O. A., Friston, K. J., and Kringelbach, M. L. (2022). Music in the brain. *Nat. Rev. Neurosci.* 23, 287–305. doi: 10.1038/s41583-022-00578-5

Weinberg, G., Raman, A., and Mallikarjuna, T. (2009). "Interactive jamming with Shimon: a social robotic musician," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, CA), 233–234. doi: 10.1145/1514095.1514152

Witek, M. A., Clarke, E. F., Wallentin, M., Kringelbach, M. L., and Vuust, P. (2014). Syncopation, body-movement and pleasure in groove music. *PLoS ONE* 9:e94446. doi: 10.1371/journal.pone.0094446

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Novelli and Proksch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fnbot.2022.844773

# The Problem of Meaning: The Free Energy Principle and Artificial Agency

*Julian Kiverstein*[1,2]*, *Michael D. Kirchhoff*[3] and *Tom Froese*[4]

[1] Academic Medical Center, Amsterdam, Netherlands, [2] Amsterdam Brain and Cognition, University of Amsterdam, Amsterdam, Netherlands, [3] Faculty of Arts, Social Sciences, and Humanities, School of Liberal Arts, University of Wollongong, Wollongong, NSW, Australia, [4] Embodied Cognitive Science Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

Biological agents can act in ways that express a sensitivity to context-dependent relevance. So far it has proven difficult to engineer this capacity for context-dependent sensitivity to relevance in artificial agents. We give this problem the label the "problem of meaning". The problem of meaning could be circumvented if artificial intelligence researchers were to design agents based on the assumption of the continuity of life and mind. In this paper, we focus on the proposal made by enactive cognitive scientists to design artificial agents that possess sensorimotor autonomy—stable, self-sustaining patterns of sensorimotor interaction that can ground values, norms and goals necessary for encountering a meaningful environment. More specifically, we consider whether the Free Energy Principle (FEP) can provide formal tools for modeling sensorimotor autonomy. There is currently no consensus on how to understand the relationship between enactive cognitive science and the FEP. However, a number of recent papers have argued that the two frameworks are fundamentally incompatible. Some argue that biological systems exhibit historical path-dependent learning that is absent from systems that minimize free energy. Others have argued that a free energy minimizing system would fail to satisfy a key condition for sensorimotor agency referred to as "interactional asymmetry". These critics question the claim we defend in this paper that the FEP can be used to formally model autonomy and adaptivity. We will argue it is too soon to conclude that the two frameworks are incompatible. There are undeniable conceptual differences between the two frameworks but in our view each has something important and necessary to offer. The FEP needs enactive cognitive science for the solution it provides to the problem of meaning. Enactive cognitive science needs the FEP to formally model the properties it argues to be constitutive of agency. Our conclusion will be that active inference models based on the FEP provides a way by which scientists can think about how to address the problems of engineering autonomy and adaptivity in artificial agents in formal terms. In the end engaging more closely with this formalism and its further developments will benefit those working within the enactive framework.

**Keywords: artificial agency, sensorimotor autonomy, the free energy principle, active inference, problem of meaning, frame problem, relevance problem**

OPEN ACCESS

***Edited by:***
*Adam Safron,*
*Johns Hopkins Medicine,*
*United States*

***Reviewed by:***
*John Alexander Vervaeke,*
*University of Toronto, Canada*
*Miriam Kyselo,*
*Norwegian University of Science and Technology, Norway*

***Correspondence:**
*Julian Kiverstein*
*j.d.kiverstein@amc.uva.nl*

***Received:*** *28 December 2021*
***Accepted:*** *17 May 2022*
***Published:*** *23 June 2022*

***Citation:***
*Kiverstein J, Kirchhoff MD and Froese T (2022) The Problem of Meaning: The Free Energy Principle and Artificial Agency. Front. Neurorobot. 16:844773. doi: 10.3389/fnbot.2022.844773*

## INTRODUCTION

The problem of meaning has haunted artificial intelligence (AI) more or less from its inception, and it still hasn't been solved. It goes by a variety of names: the symbol grounding problem; the frame problem; and the relevance problem, and it stands behind John Searle's famous Chinese room thought experiment (Searle, 1980). In what follows we will take the problem to be how to engineer artificial agents that are the source of their own values, needs and goals. Such an agent will have its own perspective relative to which its engagements with the world are imbued with meaning.

We take as our starting point Froese and Ziemke's (2009) biological principles for the design of artificial agents. First they argued for a shift in the design process toward engineering the appropriate conditions for an agent to self-generate and sustain its own identity as an individual agent under precarious conditions—a property we refer to as "autonomy" (Thompson, 2007; Di Paolo and Thompson, 2014; Di Paolo et al., 2017). Autonomy is a property of the organization of living systems that is introduced to explain how such systems can be self-individuating. Biological systems possess autonomy when the processes that make up the system form an "operationally closed" set of mutually enabling relations. The organization of the system as a whole is constantly regenerated by the activities of its constituent processes. In the absence of any of the co-enabling relations among its constituent processes, the organization of the system would break down, and is therefore described as "precarious".

The second design principle they proposed is that artificial agents should exhibit "adaptivity": the process by which an autonomous system regulates its interaction with the environment so as to avoid situations that would lead to a loss of viability, were they to be encountered. Froese and Ziemke argued that an agent that exhibits this dual profile of autonomy and adaptivity would have its own point of view on the world. Relative to this point of view, actions can be evaluated as good or bad, adequate or inadequate, successful or unsuccessful for maintaining the organism's viability.

In practice, it has proven difficult to design artificial agents that satisfy the first condition of being physically self-individuating. An alternative strategy, first proposed by Di Paolo (2003), has therefore been to design agents that acquire regular, relatively stable, and self-sustaining patterns of sensorimotor engagement with their environment (Egbert and Barandiaran, 2014; Di Paolo et al., 2017; Ramírez-Vizcaya and Froese, 2020). Instead of building robots that instantiate metabolic processes that self-organize to form autonomous networks, the strategy has been to build robots whose *sensorimotor* processes self-organize to form autonomous networks. Such stable, self-sustaining patterns of sensorimotor interaction, are the basis for what we will call "sensorimotor autonomy". The organization of sensorimotor behavior can ground the values, norms and goals necessary for an artificial agent to encounter a meaningful environment in much the same way as biological autonomy does in living systems.

It is this strategy for solving the problem of meaning in artificial agents that we take up in this paper. We will consider whether the Free Energy Principle (FEP) might provide formal tools for modeling the conditions required for an agent to acquire sensorimotor autonomy. The FEP states that organisms act to keep themselves in their expected phenotypic and ontogenetic states, and they achieve this goal by minimizing an information-theoretic quantity referred to as "free energy". In this specific sense, the FEP implies that all living systems can be modeled as if they visit a bounded and limited set of states (but not necessarily the exact same states) if they are to continue to exist (Friston, 2019). Active inference describes the process of selecting actions that minimize free energy over time. Could active inference models based on the FEP be used to mathematically model sensorimotor autonomy?

We argue first that the FEP can be applied to many systems that do not satisfy the conditions for sensorimotor autonomy, such as swinging pendulums and Watt governors (Kirchhoff and Froese, 2017; Kirchhoff et al., 2018; Baltieri et al., 2020). One can model such systems as inferring the hidden states of their observations, and thereby treat them as if they were engaged in updating their posterior distributions in accordance with Bayesian inference. We go on to distinguish physical systems like synchronizing pendulums that can be modeled as engaging in "mere" active inference from systems that are modeled as engaging in what we will call "adaptive active inference". Adaptive active inference refers to the process of actively selecting actions that minimize *expected* free energy associated with their future states (Kirchhoff et al., 2018). Mere active inference allows one to give a description of coupled systems (e.g., swinging pendulums) as inferring the hidden states of one another, thus updating their posterior beliefs. However, this is only a description. Moreover, in mere active inference, the relevant systems cannot actively change their relation to their environment. It is good to be able to update one's beliefs about the world; it is even better to be able to actively change one's relation to one's environment. It is this crucial latter aspect that is captured by shifting from mere active inference to adaptive active inference. As an example of a model of adaptive active inference, we describe a recent simulation of bacterial chemotaxis (Tschantz et al., 2020). Chemotaxis is often given as a flagship example of adaptivity. Tschantz et al. showed how their simulated agent could learn to engage in chemotaxis by means of processes of *expected* free-energy minimization. We go on to argue that adaptive active inference may well provide formal tools for modeling sensorimotor autonomy (drawing on previous work by Kirchhoff et al., 2018; Ramstead et al., 2021; van Es and Kirchhoff, 2021).[1]

Our aim in this paper is to argue that the FEP could potentially serve as a modeling technique for designing artificial agents

[1]Note that it remains an open question what the precise relation is between the simulation-based chemical agent and actual chemical agents. There are good reasons to think that in models of active inference agents, the similarity is sufficiently tight to infer properties about actual chemical agents on the basis of simulation-based models of chemical agents (see Kirchhoff et al., 2022 for further defense of this claim).

in accordance with enactive principles. We seek to use the FEP to provide enactive cognitive science with formal tools for modeling sensorimotor autonomy. Such a research programme must however confront a number of significant challenges that have emerged in the recent literature. We take up two in what follows.

First, it has recently been argued that biological systems are not well described as state-determined systems that over time are attracted toward non-equilibrium steady-states (Froese and Taguchi, 2019; Aguilera et al., 2021; Di Paolo et al., 2022). These authors have argued that organisms (perhaps in contrast to FEP-based models of agency) have a natural history that is characterized by open-ended, unpredictable transitions to qualitatively new regimes of order. Di Paolo et al. (2022, p. 21) give as examples "embryogenesis, life-cycle patterns, epigenetic variability, metamorphosis, and symbiosis." They argue furthermore that qualitative transformations can be observed in the structure of behavior in the learning of skills, and in the soft assembly of task-specific systems in tool use (Anderson et al., 2012; Di Paolo et al., 2017). These critics have argued that processes of historical change are essential to adaptivity but such history-dependent processes cannot be captured in the terms of the FEP. Once a system returns to a non-equilibrium steady-state its history is forgotten. If these critics are correct, there are therefore essential differences between systems that engage in adaptive active inference, and biological agents that exhibit sensorimotor autonomy.

Second, Aguilera et al. (2021) have argued that a free energy minimizing system would fail to satisfy a key condition for sensorimotor agency referred to as "interactional asymmetry". They show how the mathematical assumptions the FEP rests upon only apply to systems whose interactions with the environment are symmetrical. If Aguilera et al. are correct, the mathematics of the FEP is not well suited for modeling sensorimotor autonomy. The FEP doesn't take us any further forward in understanding the formal properties of systems that are the source of their own values, needs and goals.

We finish up by offering reasons why the door should remain open to a synthesis of the FEP and enactive cognitive science we propose in our paper. First, we argue that the FEP is highly general, applying to both systems that implement mere active inference as well as to systems that are able to perform adaptive active inference. We suggest this generality is an advantage of the FEP allowing it to approximately represent a wide range of different systems including, if the arguments of our paper hold up, systems that fall in the region of those possessing sensorimotor autonomy. Second, we will argue that systems that implement adaptive active inference will tend to exhibit transient or metastable dynamics in which there is a recurring creation and destruction of large-scale coordination dynamics. Although metastable systems can be described as on average revisiting their attracting states they will avoid ever settling into any of these attracting states. Metastable systems exhibit the kind of historical, path-dependent learning required for acquiring a sensorimotor identity, and becoming an agent. Thus the key question for the FEP is whether adaptive active inference can be used to model systems with metastable dynamics. We will provide reasons for returning a positive answer to this question; though the work of building such formal models, so far as we know, remains to be done.

We conclude that the two frameworks need each other. Enactive cognitive science needs the FEP to formally model the properties it argues to be constitutive of agency. The FEP needs enactive cognitive science for the solution it provides to the problem of meaning. In the end engaging more closely with this formalism and its further developments will benefit those working within the enactive framework.

## The Enactive Approach to the Problem of Meaning in Artificial Intelligence

Biological agents are able to act in ways that express a sensitivity to context-dependent relevance. Organisms engage with an environment that is structured by their practical involvements, cares and concerns. Minimally, organisms have a concern for their own continued existence and their manners of living. Organisms must for instance engage in a continuous struggle to stave off death. Human agents are of course concerned with much more than meeting basic biological needs required for survival. Their activities are animated and driven by a variety of desires they strive to satisfy, many of which stem from distinctively human, sociocultural ways of living.

The problem of meaning arose in artificial intelligence in attempting to design artificial agents that are able to act adaptively and flexibly in dynamic complex and open-ended real-world situations. A popular approach in artificial intelligence research has been to build systems that learn an internal model of their environment and that make inferences and plans on the basis of this internal model (e.g., Lake et al., 2017). The sensitivity to what is relevant in a perceived situation has however proved resistant to specification in ways that could allow for this sensitivity to be captured in an internal model. To act adaptively and flexibly in dynamic complex environments such a system will need to determine from its internal model what is actually relevant under conditions of continuous change. Everything the system knows is of possible relevance. How then does the system determine what is of actual relevance without engaging in an exhaustive search of everything it knows (Dennett, 1984; Dreyfus, 1992; Fodor, 2000; Wheeler, 2005; Samuels, 2010; Vervaeke et al., 2012; Vervaeke and Ferraro, 2013; Danks, 2014; Shanahan, 2016)? The problem could perhaps be solved if artificial intelligence researchers could find a way to make an internal model that represents all possible contexts in terms of their determinate properties. However, the environment that living beings encounter in perception is not an environment made up of objects and properties that stand in determinate logical relations. As we began this section by noting, organisms perceive an environment that is structured by their needs, cares and concerns.

In what follows we will take the problem of meaning to be equivalent to what is sometimes called the "relevance problem." Thus, we will use the terms "meaning" and "relevance" interchangeably in what follows. Meaning and relevance should be distinguished from information which we take to refer to statistical correlation between states of two systems (e.g., an

organism and its environment). We take it to be uncontroversial that statistical correlation does not suffice to make it the case that the states of a system are meaningful for the system (Hutto and Myin, 2013). In living systems what makes a state the bearer of meaning is the history, dynamics and structure of the system (Varela, 1979; Oyama, 2000; Thompson, 2007). The history and structure of the living organism serve as the basis for needs, goals and values that move the organism to act in its environment. Meaning is determined by the organism's history, dynamics and structure. We identify meaning with relevance because we take meaning to be brought forth by the agent through a history of engagement with an environment that is relevant to the agent because it is structured by the agent's needs, concerns and values.

Froese and Ziemke (2009) argued that the problem of meaning could be circumvented if artificial intelligence researchers were to design agents based on the assumption of the continuity of life and mind. The core idea behind life-mind continuity is that intelligence depends upon its biological embodiment, where embodiment is to be understood in terms of the organizational properties of autonomy and adaptivity. The idea of continuity is therefore that the concepts and principles required for understanding and explaining features of mind such as subjectivity, agency and individual identity, are also the principles and concepts employed to explain the phenomenon of life (Kirchhoff and Froese, 2017; Di Paolo, 2018: p.74). Agents that are biologically embodied are the source of their own norms, values and goals. They escape the problem of working out from everything they know, what is relevant to their current and future contexts of activity. Relevance is not an extra ingredient that has to be added to what the agent already knows but is instead intrinsic to what is perceived. To see how this follows, we must further unpack the key concepts of biological autonomy and adaptivity briefly encountered in our introduction.

To possess biological autonomy a system must first of all be operationally closed. That is to say it must be organized so as to produce "a network of precarious processes in which each process enables at least one other process in the system and is, in turn, enabled by at least one other process in the system" (Di Paolo et al., 2017: p.113). The operationally closed network has a precarious existence insofar as the constituent parts that make up the network as a whole, are processes that stand in co-enabling relations. Each process would break down were it not causally enabled by the other processes in the network. The component processes are co-enabling insofar as they work together to produce the larger network as a whole. The self-production of the network as a whole is a task that needs to be continually accomplished if decay, disintegration and death are to be avoided. The system can therefore be said to be biologically autonomous in the sense that it is the operation of the processes that make up the system that enable its continued self-production and its self-distinction from its surroundings.

Systems that are biologically autonomous constitute or produce themselves as individuals – they are self-individuating.[2] This process of self-individuation serves as the basis for agency – the organism is able to distinguish and actively regulate flows of energy and matter that contribute positively to its self-individuation, and avoid those that potentially interfere with its biological autonomy (Varela, 1991; Di Paolo, 2005; Thompson, 2007). The organism's coupling with its environment is inherently risky because of the precariousness of the processes that produce and sustain the organism's continued existence. To succeed in its goal of continuously realizing processes of self-production, the organism must be selectively open to energetic exchanges with the environment that contribute to the conditions of its self-production, and closed to exchanges that threaten its self-distinction (Di Paolo, 2018: p.84). Autonomy thus underwrites a basic biological form of normativity – the capacity to differentiate between, and thereby regulate, flows of matter and energy based on how well or badly they contribute to the organism's goal of maintaining its precarious identity. The classic example of biological normativity, and one we will return to later, is chemotactic behavior in which a bacterium will move toward metabolisable compounds and move away from metabolic inhibitors.

The biological normativity that is intrinsic to autonomy is not dependent on the observer's perspective on the organism's behavior. It is a capacity that is intrinsic to the organism's biological organization. The organism's capacity to regulate and modulate its relation to the environment is dependent on a sensitivity to dynamical trajectories, gradients, and tendencies Di Paolo has labeled "adaptivity". Adaptivity is agentive in the sense that it is a capacity the system actively exercises in changing the parameters and conditions of the agent-environment relation in for instance seeking out food when energy is anticipated to be needed. This active modulation introduces an asymmetry into the organism's coupling with its environment, referred to in the literature as "interactional asymmetry" (Barandiaran et al., 2009; Di Paolo et al., 2017, §5.2.2). The organism modulates its relation to the environment based on its sense of whether environmental events are good or bad for its continued existence.

Di Paolo (2005) defines adaptivity as:

"A system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,

1. Tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,
2. Tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity." (Di Paolo, 2005: p. 438)

The reason this is important is because it implies that the organism need not passively respond to environmental events in a state-determined manner based only on its previous state. The organism's operating conditions can undergo change over

[2]Maturana and Varela's (1980) concept of autopoiesis is a paradigm example of biological autonomy. An autopoietic system is composed of co-enabling processes that form an operationally closed network, and that dynamically produce the system's own material boundary. It should be noted that systems can be autonomous while lacking a material boundary. This is the case for insect colonies for instance that form autonomous social networks (Thompson, 2007: p.44).

time based instead upon its history of interactions with its environment. The organism can in this way have a plastic identity that is given shape by its history of acting, and being acted upon by its environment. Meaning can thus be understood as actively generated or brought forth by the organism based on the history of sensorimotor interaction with its environment that has become sedimented in its biological organization. So conceived, meaning does not need to be somehow added to what the organism knows because the environment the organism relates to is always already imbued with meaning based on the organism's past history of interaction.[3]

So far our treatment of the concepts of biological autonomy and adaptivity has focused on processes that the organism depends upon for its continued viability. However, the norms relative to which the organism regulates its interactions with the environment do not only concern its continued existence in the here and now. The significance of sensory perturbations for the organism go beyond their immediate bearing upon the organism's operationally closed organization.[4] The processes that constitute and produce the organism as an *agent*, include its regular and relatively stable patterns of sensorimotor behavior. These patterns of sensorimotor behavior have been argued to also exhibit the key properties of autonomy—they depend upon operationally closed sensorimotor networks made up of co-enabling bodily and environmental processes (Di Paolo, 2005; Barandiaran, 2008; Egbert and Barandiaran, 2014; Di Paolo et al., 2017; Ramírez-Vizcaya and Froese, 2020). Think for example of habits like smoking cigarettes or drinking coffee when you wake up in the morning. These are sensorimotor patterns of behavior that are self-sustaining, but that do not positively contribute to maintaining the organism's biological viability, and may even be harmful to the organism. A pattern of behavior becomes a self-sustaining habit when the processes that enable it—neural, muscular and environmental—depend for their stability and organization on the regular performance or enactment of the pattern of behavior (Di Paolo et al., 2017: p.144; also see Egbert and Barandiaran, 2014). Thus these processes come to form operationally closed sensorimotor networks in much the same way as metabolic processes do. At the same time, the organization of the sensorimotor network is precarious because it is at risk of extinction if the pattern of behavior is not regularly enacted.

It has recently been proposed that sensorimotor autonomy could serve as a design principle for artificial agents that would allow researchers to avoid the difficult problem of engineering systems that metabolically self-produce. Di Paolo (2003) suggested for instance that robots could be built with mechanisms "for acquiring a way of life, that is, with habits" (p.31). Designing agents that can acquire self-sustaining habits will have the consequence that such agents will engage with the world based on norms, goals and values that relate to the sustaining of their habits. They will differentially evaluate the situations they encounter in terms of their relevance for the realization of processes upon which the sustaining of their habits depend. Such an agent doesn't relate to an action-neutral world that stands in need of representation. It will not need to work out from all possible responses, which responses are actually relevant to its current situation. Instead agent and environment will form a single system that is continuously reconfigured in ways that allow for the sustaining of the sensorimotor autonomy of the agent. This is, in a nutshell, the enactive proposal for how to solve the problem of meaning in artificial intelligence.

Still a question remains of how to model sensorimotor autonomy. The question we take up in the rest of our paper is: could the free energy principle (FEP) provide a formal description of the conditions for the design of an artificial agent that possesses sensorimotor autonomy?

## The Free Energy Principle: A Brief Introduction

The FEP purports to describe the organizational properties a system must instantiate if it is to preserve its organization over time in its interaction with a dynamic environment. The FEP has been argued to apply to "any biological system…from single-celled organisms to social networks" Friston and Stephan (2007). It claims that all complex adaptive systems that are able to resist a tendency to disorder must minimize an information-theoretic quantity known as "free energy". Friston (2010) for instance formulates the FEP as follows:

> "The free-energy principle… says that any self organizing system that is at equilibrium with its environment must minimize its free energy. The principle is essentially a mathematical formulation of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to disorder." (Friston, 2010: p.127).

The FEP is sometimes described as a tool the scientist employs purely for modeling purposes. Raja et al. (2021) for instance formulate the FEP as claiming: "Any ergodic random dynamical system with an attractor and a Markov blanket behaves *as if* it were minimizing the variational free energy of its particular states" (p.3, our emphasis). The "as if" qualifier here is used to indicate that the behavior of complex adaptive systems is modeled on the assumption that adaptive systems minimize variational free energy. It doesn't matter for modeling purposes if this assumption is true. A number of papers argue on this basis that strictly speaking the FEP has nothing to say about the organizational properties of the complex adaptive systems it purports to model (Ramstead et al., 2020b; van Es and Hipólito, 2020; van Es, 2021). These authors argue the FEP should be understood in purely instrumental terms as a scientific tool for predicting the observable behavior of adaptive systems. Our paper is premised on the assumption that such an instrumentalist

[3]Von Uexküll (1957) introduced the idea of a sensorimotor world (an *Umwelt*) to characterize this meaningful environment. The *Umwelt* is the world as it is perceived by an organism given its sensorimotor repertoire.

[4]Di Paolo et al. (2017) refer to Margaret Donaldson's distinction between four modes of human thinking about the world, which she distinguished based on the degree of decentering from the agent's point of view on the world in the here and now. Humans can for instance plan for future events, and they can detach themselves entirely from their own concerns in thinking about moral and political values. Di Paolo et al. appeal to what they call "virtual actions" to account for the development of these decentered modes of thinking out of sensorimotor agency (see also Kiverstein and Rietveld, 2018). It is beyond the scope of our paper to discuss their account of virtual actions in further detail.

reading of the FEP is incorrect [for further discussion see Andrews, 2021; Kirchhoff et al., 2022; Kiverstein and Kirchhoff, 2022]. Our aim in this paper is to consider whether the FEP can be used to formally model sensorimotor autonomy. We take sensorimotor autonomy to be a real organizational property that tells us what it is for a system to be an agent. Our aim is to consider if models based on the FEP can be taken to truthfully represent a real organizational property of agents.

The FEP, as we will understand it, employs the mathematical formalism of non-equilibrium steady-state (NESS) systems to model the properties a complex adaptive system must instantiate if it is to preserve its organization over time (Friston, 2012, 2013, 2019). Any biological system will be able to maintain order within a boundary (modeled as a "Markov blanket," more on which below), separating the internal states of this system from the external states of its environment. The FEP claims that to maintain order within this boundary, the system must (on average, and over time) revisit a set of sensory states when it is perturbed by the environment. We will refer to the set of sensory states that the system is modeled as repeatedly revisiting over time as the "attracting set" for a given biological system. We can think of the attracting set as a model of the system's extended phenotype since it will include variables for morphological states as well as behavioral patterns that relate to the niche the agent constructs (Friston, 2011; Kirchhoff and Froese, 2017; Bruineberg et al., 2018; Kirchhoff and Kiverstein, 2019). A system's attracting set will include physiological states such as blood oxygen concentration and pressure levels and body temperature that must be maintained within a certain range of values if the organism is to survive. Other sensory states belonging to a system's attracting set relate to its niche - fish frequent aquatic environments, while humans tend to live on land and only occasionally find themselves underwater. The states belonging to the system's attracting set will therefore be the subset of all possible states the system can occupy that are highly probable given the system's phenotype and the niche it inhabits. States that fall outside the attracting set are potentially threatening to the maintenance of order within the system because they lead to an increase in disorder or entropy within the system. States that lead to an increase in disorder will be surprising or improbable for a NESS system that tends toward an ordered set of states over time in its exchanges with the environment. The states belonging to the system's attracting set are states the system expects to occupy over time. When the states of the system fall outside of its attracting set this is therefore surprising because the probability of finding the system in such states is low. ("Surprise" is to be understood as the improbability of a particular sensory state, and is not to be confused with agent-level surprise, which occurs in response to an unexpected conscious sensation).

The system has no tractable way of calculating whether a given sensory state is surprising or not. This is because the probability of a sensory state is calculated relative to a state of the possible influences of external states of the environment on the internal states of the system. The state space is however potentially infinite, thus computing the probability of each sensory state by searching through this state space will prove intractable. This is where free energy can help, since free energy is a quantity that can act as an upper bound on surprise. Free energy more technically is a function of the function of sensory states that is parameterized by the internal states of a system. Since free energy is a function of the sensory and internal states of the system, it is in principle computable (Friston and Ao, 2012). Moreover, it is a quantity over which the organism has (indirect) control since it maps onto the organism's sensory states that it can control through action, and internal states that admit of a certain degree of plastic reorganization through learning. Minimizing free energy will guarantee that sensory states remain in a high-probability area in the system's state space. So long as the NESS system can keep the free energy associated with its sensory states to a minimum, it will succeed in remaining in states that belong to its attracting set.

The FEP states that all quantities that can change in a NESS system will change to minimize free energy (Friston and Stephan, 2007). Free energy quantifies the mismatch between the sensory states the system *expects* to sample through its actions, and those it actually samples. The notion of "expectation" should be understood in relation to a model that is entailed by the internal dynamics that form in the system's coupling with its niche. The function of this model is to anticipate sensory perturbations originating in the environment external to the system, allowing the system to proactively adapt its actions to those perturbations.

The FEP models complex adaptive systems as random dynamical systems that are attracted toward a non-equilibrium steady-state (a NESS). The FEP assumes adaptive systems will tend to exhibit certain dynamical flows of states over time determined by, amongst other things, their phenotypic states, body morphology, and their ecological niche. Generative models are used to describe the statistics of these flows (Ramstead et al., 2020a). For a system to tend to flow toward a NESS by minimizing free energy is for the system to minimize the discrepancy between the variational density (also sometimes called the "recognitional density") the organism instantiates in its internal dynamics, and the true posterior or the external dynamics in the environment. Free energy can thus be thought of as quantifying mathematically the mismatch between the organism's internal dynamics and the external dynamics of its environment (Bruineberg et al., 2018).

Friston (2013) has proposed that a living system does not have a model of its environment but it is a model of its environment, which highlights that the notion of "model" the FEP is premised upon is implicit in the living system's internal dynamics. In this sense, there is no distinct system inside of the central nervous system of the agent that uses a model to engage in inference. For Friston, inference just is a description of the flow of the internal dynamics of the living system. Friston takes the generative model to be organized around the organism's belief in its own continued existence. All of the actions the organism undertakes aim at sampling sensory states that maximize the evidence for this belief in its continued existence, a belief Allen and Tsakiris (2018) have referred to as the "first prior". To minimize free energy is at one and the same time to maximize evidence for this belief in the living system's continued existence. Hohwy (2016) refers to this property of living systems whereby they act to sample

evidence that confirms the belief in their own continued existence as "self-evidencing".

Free energy can be minimized in two intimately related ways referred to as "perceptual" and "active inference". In perceptual inference free energy is reduced by changing the dynamics internal to the system (Friston, 2010; Hohwy, 2013). The internal dynamics of the system embody a model of the agent's econiche by means of which it can steer its actions (Friston, 2011). Perceptual inference involves plastically restructuring the internal dynamics in such a way that the agent is better able to accommodate external sensory perturbations arising from the changes in its niche in the future (Friston et al., 2016). Free energy is kept to a minimum in part by generating and modifying an internal dynamics that closely approximates the external environmental dynamics. We said that perceptual and active inference are intimately related (Hohwy, 2013). This intimate relation follows from what we have just referred to as self-evidencing (Hohwy, 2016): the internal dynamics that are adjusted in perceptual inference are organized around sampling sensory evidence that confirms the agent's belief in its own continued existence (Fotopoulou and Tsakiris, 2017; Allen and Tsakiris, 2018; Seth and Tsakiris, 2018).

In active inference the agent acts to sample sensory states belonging to its attracting set (Friston et al., 2017a,b). The sensory states that are expected given the first-prior are those that relate to the agent's needs, goals and intentions (Allen and Tsakiris, 2018). The agent's continued existence will for example depend on its meeting its biological needs for warmth, nourishment, and attachment (Fotopoulou and Tsakiris, 2017). If the agent is to sample sensory states that maximize the evidence for the first prior, this will require the agent to act in ways that satisfy such basic needs. A simple example is eating when hungry. Hunger indicates a potential breach of essential variables relating to blood glucose levels (i.e., a deviation from the system's attracting set). The action of eating helps to correct this potential breach before it arises. With this brief summary in place we turn next to the question of whether the FEP provides a formal description of the conditions required for an artificial agent to possess sensorimotor autonomy.

## The Free Energy Principle: A Minimal Condition for Sensorimotor Autonomy?

Recall our proposal is to use sensorimotor autonomy as a biologically-based design principle for building artificial agents (Barandiaran, 2008; Egbert and Barandiaran, 2014; Di Paolo et al., 2017; Ramírez-Vizcaya and Froese, 2020). The idea is that habits are self-sustaining patterns of activity that constitute a systemic identity for the agent relative to a sensorimotor domain. Relevance arises out of the needs, goals and interests the agent has in sustaining its habits. Situations and activities "become meaningful not only in virtue of their contribution to biological survival, but also in virtue of their contribution to the stability and coherence of a sensorimotor repertoire" (Di Paolo et al., 2017, p.39). An agent that has sensorimotor autonomy will have its own point of view relative to which evaluations of action possibilities can be made in terms of their relevance for the agent. Does the FEP provide a set of mathematical tools that can be used to model sensorimotor autonomy? Is free energy minimization sufficient for sensorimotor autonomy? The payoff for a positive answer to this question will be formal tools that allow us to connect meaning and relevance to a system's intrinsic dynamics.[5]

The FEP is broad in terms of the systems to which it applies. Swinging pendulums, Watt governors and pebbles have all been argued to count as systems that can be described as minimizing free energy in their dynamic coupling with the environment (Kirchhoff and Froese, 2017; Kirchhoff et al., 2018; Baltieri et al., 2020; van Es and Kirchhoff, 2021). Two coupled pendulums A and B can, for example, be described as modeling each other's motion. Given the internal states of pendulum A, and the effects of its velocity and motion on the beam from which it is hanging, pendulum A can be said to infer the motion of pendulum B. This is possible because the motion of pendulum A, through its effects on the beam, enslaves the motion of pendulum B, and vice versa. When the two pendulums come to swing in synchrony the coupling of the two pendulums can therefore be described in terms of free energy minimization (Bruineberg et al., 2018; Kirchhoff et al., 2018). In line with our earlier work, we describe this process of free energy minimization that can be observed in non-living, and non-cognitive systems as "mere active inference" (Kirchhoff et al., 2018). Each pendulum infers through its own motion and the effects of its motion on the beam, the motion of the other pendulum.

Mere active inference is qualitatively different from the process of free energy minimization that occurs in living and cognitive systems (Kirchhoff, 2018). Living systems are able to sample among different options, and select the option that has the least *expected* free energy.[6] While the pendulums enslave each other's motion, living systems are able to free themselves from their proximal conditions by selecting temporally extended sequences of actions that minimize expected free energy associated with future states. We have used the term "adaptive active inference" to describe what living systems are able to do that is missing in systems that engage only in mere active inference (Kirchhoff et al., 2018). In adaptive active inference sequences of actions are selected that minimize the cumulative

[5]Kolchinsky and Wolpert (2018) have recently explored a strikingly similar proposal. They begin their paper by noting the difference between what they call "semantic information" and Shannon information as a measure of statistical correlation between two systems. They seek to provide formal tools that describe how semantic information could be intrinsic to the dynamics of a system in a given environment. Central to their proposal is a viability function which they define as "the negative Shannon entropy of the distribution over the states of the system X" (p.2). They use Shannon entropy as an upper bound on the probability that the system occupies states belonging to its viability set in a similar fashion to how Friston uses variational free energy in relation to a system's NESS. An important difference is that Shannon entropy is not computationally tractable for the agent, whereas variational free energy represents a quantity an agent can control through its actions. We will postpone exploring further the similarities and differences in our proposals. What we wish to emphasize for now is that both our proposals aim to formalize meaning (Kolchinsky and Wolpert use the term "semantic information") in terms of a quantitative measure of viability - in our case variational free energy.

[6]The distinction between mere and adaptive active inference is formally grounded. For discussion see Millidge et al. (2021), though they do not use our terminology of "mere" and "adaptive" active inference.

sum of free energy over time, a quantity referred to as "expected free energy" (Friston et al., 2017a,b).

Adaptive active inference is distinguished from mere active inference in aiming at the selection of actions whose sensory effects minimize expected free energy. (Expected free energy is the free energy expected upon executing a temporally-extended sequence of actions.) Expected free energy is a function of two quantities referred to as instrumental and epistemic value.[7] To minimize expected free energy an agent must select action policies (sequences of actions) that maximize both instrumental and epistemic value. Instrumental value is maximized when the sensory observations an agent expects to sample match its preferred outcomes (its needs, goals and desires). Thus, acting to maximize instrumental value can be thought of as equivalent to goal-directed behavior. Epistemic value quantifies information gain or the reduction of uncertainty about the hidden states of the environment. An agent maximizes epistemic value by maximizing the information that is gained through exploratory actions of the environment. An active inference agent that acts to minimize expected free energy will continuously be balancing instrumental actions that aim at bringing about preferred outcomes with epistemic actions that aim at uncertainty reduction. Crucially, while this kind of epistemic (or information seeking) foraging should *on average* result in the minimization of uncertainty, there will nevertheless be short-term peaks of uncertainty given an organism's exploration of its surroundings. The aim is thus to strike the right balance between the reduction of entropy and temporarily increasing entropy. The pay-off for finding this right balance (what is sometimes called the "exploitation-exploration trade-off") is that the agent will avoid getting trapped in any local minima. They will be able to make continuous progress and improvements in learning in ways that are conducive to long-term free energy minimization. (For further discussion see Kiverstein et al., 2019).

There are other points of importance to note about adaptive active inference. First, the generative model is biased toward sampling sensory observations that match the agent's preferences, goals and desires (Bruineberg et al., 2018; Tschantz et al., 2020). Second, and relatedly, epistemic actions will work in the service of tinkering with a model that is biased toward the control of certain sensory outcomes. As Tschantz et al. have noted, an active inference agent will tend to forage for information in parts of the environment expected to maximize instrumental value (Tschantz et al., 2020: p.7). That is to say, the improvement in the model that epistemic actions make possible are ultimately improvements in the service of the agent's goals.

To minimize expected free energy the agent has to select from among action policies, the policy that is expected to lead to preferred outcomes and goals (Friston et al., 2017a,b; Pezzulo et al., 2018). This might be thought to lead the active inference agent to encounter the relevance problem once again.[8] The agent will always be faced with an open-ended range of possible action policies but can only search a narrow area within this space. How then does the agent constrain the search space to only action policies of relevance (i.e., those expected to minimize free energy? Most active inference models up until now have avoided this question by pre-specifying the search space. Within this predefined search space action policies are then selected on the basis of the agent's belief in the precision of the policy - the confidence the agent places in the sensory consequences of its actions. The work of scoring action policies is taken over by the precision estimate associated with each action policy. Precision estimates are based on expected uncertainty (or salience) and unexpected uncertainty (or volatility, Parr and Friston, 2017). The higher the precision for each action policy, the more confident the agent can be that the sensory outcomes of its action will match its preferred outcomes. A risky action policy is one whose sensory consequences the organism anticipates will diverge from its preferred outcomes leading to an increase in expected free energy. Precision estimates can be thought of as having effects comparable to attention. They bias action selection toward actions whose sensory consequences are expected to minimize free energy. The "gain" is turned-up on opportunities to bring about those sensory consequences. Precision is decreased and the gain turned-down on actions whose sensory consequences are associated with increases in free energy.

Is there reason to believe that adaptive active inference will scale-up from a predefined search space of action policies, without the agent once again encountering the relevance problem?[9] Recall how we are proposing that artificial agents that develop sensorimotor autonomy will circumvent the problem of meaning. Meaning will arise out of the agent's history of activity in an environment structured by its needs, interests and concerns. Meaning is not an extra ingredient the agent needs to add to information to determine how to solve what would otherwise be an ill-defined problem. "With ill-defined problems, the goal-state is often murky, the initial state is unclear," and the operations that will take you from your initial state to your goal state are unspecified (Vervaeke and Ferraro, 2013, p.4). Before one can solve an ill-defined problem one must determine what information is relevant for defining the problem. Our hypothesis is that agents that possess sensorimotor autonomy however will typically not encounter ill-defined problems.[10] They will relate to an environment that is already meaningful because of their past

[7]For mathematical details of how epistemic and instrumental value are computed we refer the reader to Tschantz et al. (2020, pp.25–26), and Friston et al. (2017a). For discussion of the relation between variational and expected free energy see Millidge et al. (2021).

[8]Our thanks to an anonymous reviewer for raising this objection.

[9]For recent reviews of the application of active inference models in robotics see Lanollis et al. (2021) and Da Costa et al. (2022). We discovered these papers only after completing the writing of this manuscript and plan to discuss them fully in follow-up work.

[10]We do not claim that skilled agents never encounter ill-defined problems. We suggest exploratory or "epistemic" actions that aim at uncertainty reduction will provide an important part of the answer to how agents solve such problems when they do arise (see e.g. Friston et al., 2017b). It is an important question for further research whether agents engaging in epistemic actions to solve an ill-defined problem would once again encounter the relevance problem. How is it that agents performing epistemic actions to solve an ill-defined problem constrain the space of possible solutions they sample? This question is partially addressed by active inference models of curiosity and insight (see e.g. Friston et al., 2017a). Thanks to one of our reviewers for raising this problem.

history of engagement. The habits they have developed provide them with know-how or skills that form the basis for norms that guide the agent's actions. Situations and activities are good or bad, adequate or inadequate, successful or unsuccessful to the extent that they contribute to the sustaining of the agent's sensorimotor identity.

We will consider next if models of adaptive active inference could be used to formally describe the organizational property of sensorimotor autonomy. To address this question we need to briefly introduce the Markov blanket formalism. The terminology of Markov blankets is borrowed from the literature on causal Bayesian networks (Pearl, 1988; Bruineberg et al., 2022). The Markov blanket for a node in a Bayes network comprises the node's parents, children and parents of its children. The behavior of the blanketed node can be predicted from the states of the blanket without knowing anything about the nodes external to the blanket that are the causes of changes internal to the network. We suggest the Markov blanket formalism can be used to model sensorimotor autonomy. Here we make the case only informally and schematically. It is a task for future research to turn our philosophical argument into concrete formal models.

Our core idea is that the autonomy of the sensorimotor network can be modeled as the nesting relations among Markov blankets in systems that perform adaptive active inference. Each component process in the system can be thought of as having its own Markov blanket. Two components A and B stand in an enabling relation when the active states of the Markov blanket of A cause the sensory states that belong to the attracting set of B (i.e., the sensory states that B must occupy if it is to remain viable). B will begin to break down when the sensory states that form its Markov blanket are improbable, departing from what is expected given its attracting set. Thus B's continued viability is enabled by the active states of A. Conversely, component B enables component A if the sensory states belonging to A's attracting set are made highly probable by B's active states. So long as the Markov blankets of each of the component processes couple in such a way that each of the components remains in high probability sensory states, (a condition that will be satisfied in systems that engage in adaptive active inference) the result will be the self-production and self-distinction of the system as a whole (Ramstead et al., 2021; van Es and Kirchhoff, 2021). A system that engages in adaptive active inference will succeed in maintaining operational closure under precarious conditions.

Nave (2022) criticizes the use of Markov blankets to model metabolic self-production. She argues that organisms are intrinsically unstable structures that define their boundaries while undergoing near constant material turnover. To deploy the Markov blanket formalism we would first need to identify the organization of the system of interest, which is a challenge in living systems undergoing continuous material change. She concludes that the Markov blanket formalism can only be successfully deployed if we already know the organization of the system we are interested in modeling. Along similar lines, Raja et al. (2021) have argued that while the cell membrane is the product of the activity of cells, the Markov blanket is not the product of the activity of a cognitive system's internal states. They conclude: "There is nothing in the use of Markov blankets that accounts for the fundamental features of the boundary of self-organized, self-maintained systems" (p.28-9; cf. Suzuki et al., 2022).

We suggest in response that the self-production of living systems is understood as an example of autonomy (i.e., the production and maintenance of an operationally closed network under precarious conditions). Such a characterization of the organization of living systems fits perfectly with Nave's description of organisms as "intrinsically unstable structures - stabilized only via their own ceaseless activity" (Nave, 2022, preprint, p.4), and with Raja et al.' concept of constitutive self-organization. We have just shown informally how the Markov blanket formalism could be applied to systems that are modeled as engaging in adaptive active inference. To repeat the main idea: the sensory states that define the Markov blanket for each component of an operationally closed system will be coupled to the active states of one or more of its enabling components. So long as the system engages in adaptive active inference this coupling relation will ensure that the sensory states for each component belong to the component's attracting set. The result will be the self-production of the system as a whole as a unity distinct from its environment.

As a proof of concept example of how adaptive active inference can be used to model sensorimotor autonomy (but not biological autonomy), consider the recent active inference simulation of chemotaxis of Tschantz et al. (2020). "Chemotaxis" refers to the running and tumbling movements bacteria exhibit when they encounter a chemical gradient that is a potential source of food (i.e., a sucrose gradient). This can be thought of as a form of pragmatic action in which the bacterium acts to maximize instrumental value. When bacteria sense a negative gradient (i.e., an acid that is toxic to the bacterium), the rhythm of the running and tumbling motions alters in such a way as to steer the bacterium away from danger, and in search of locations were positive gradients are to be found. This behavior can be thought of as an epistemic action the bacterium performs to maximize epistemic value.

Tschantz et al. simulated an active inference agent that selected between actions by seeking to maximize both instrumental and epistemic value. They showed that in their simulation agents employing such a strategy were able to perform at least some chemotaxis (i.e., running toward positive gradients, and tumbling away from negative gradients). The strategy of minimizing expected free energy seems to have allowed the active inference agent to find the right balance between performing epistemic exploratory actions of tumbling and instrumental actions of moving forward. The agent engaged in tumbling behaviors when it estimated there was less instrumental value in running. In doing so it learned about the effects of tumbling, and continued to do so until the value of tumbling becomes less than the value of running when the agent switches its behavior.

Crucially, the value the simulated agent assigned to actions was modeled by the change in free energy over time. The policy of tumbling for instance decreases in value when the agent is no longer making information gains that resolve model uncertainty, a situation that can be understood in terms of free energy remaining constant or increasing. The policy (i.e. sequences of

actions expected to minimize FE) of running takes on a value that outweighs that of tumbling when the agent expects sensory observations that match those it prefers (i.e., a positive gradient). The increased instrumental value of running can therefore be equated to an expected reduction in free energy. This is important because valence has been analyzed and modeled in the FEP literature in terms of change in free energy over time (Joffily and Coricelli, 2013; Van de Cruys, 2017; Kiverstein et al., 2019; Hesp et al., 2021). "Valence" refers to the positive or negative charge of an affective state.

The rate of change in free energy can be taken as a measure of how well or badly the organism is faring in its interactions with the world. When free energy is on the increase, or is not resolved through action, this means that the agent is in a potentially threatening situation, while when free energy is decreasing this is feedback for the agent that it is faring well and should, if possible, continue on the same path. We suggest then that Tschantz et al.' active inference agent exhibits adaptivity in its chemotactic behavior. The active inference agent uses changes in free energy to negotiate the trade-off between performing epistemic and pragmatic actions, as we have just explained. The changes in free energy over time are used by the agent as feedback that signals how well it is doing in its goal of achieving chemotaxis, and the simulated agent modulates its coupling with its environment on the basis of this feedback. In the next section we take up two objections that challenge the hypothesis we have been proposing that the process of adaptive active inference can be used to model sensorimotor autonomy.[11]

## Ergodicity, Historicity and Interactional Asymmetry

The first objection we will consider targets the ergodicity assumption that early iterations of the free energy principle relied upon (e.g., Friston, 2013). Briefly, "ergodicity" refers to "the time average of any measurable function of the system converges (almost surely) over a sufficient amount of time. This means that one can interpret the average amount of time a state is occupied as the probability of the system being in that state when observed at random." (Friston, 2013, p. 2) If ergodicity holds, the proportion of time a system spends in any region of its phase space is equivalent to the probability of the system occupying this region of its phase space. For example, if the probability of a coin landing heads is 50/50 then over the course of the time spent flipping a coin, the coin will spend 50% of this time landing heads, and 50% of this time landing tails. We can think of the average time a system spends in any region of its phase space – the space of all possible states of the system – as being proportional to the attractiveness of that region. Recall the idea of an attracting set, that living systems as random dynamic systems, will have a set of sensory states toward which they will continually evolve over time whenever they are perturbed. This idea has been taken by critics to be based on the assumption that living systems literally are ergodic.[12]

It has recently been argued that the enactive concept of adaptivity is fundamentally at odds with the ergodicity assumption (Di Paolo et al., 2022; also see Colombo and Wright, 2018; Kauffman, 2019 for a critique of ergodicity as applied to living systems). Adaptivity, they have argued, involves changes in the phase space of the dynamical system the organism forms with the environment to avert the potential loss of viability that would ensue, were the agent to remain in a steady-state regime. The possibility of such critical transitions in an organism's phase space requires an understanding of the change in internal dynamics the agent undergoes as path-dependent, that is, as dependent on the agent's history of interaction with the environment. We see examples of such phase transitions in development, in for example, "embryogenesis, life cycle patterns, epigenetic variability, metamorphosis and symbiosis" (Di Paolo et al., 2022: p.21). In behavior, critical transitions occur in perceptual learning, skill acquisition, tool use and habit formation. Over shorter time scales, changes in patterns of effective connectivity in the brain that allow for many-to-many mapping between neural structure and function, or what Anderson (2014) calls "neural reuse", depend upon such critical transitions. In short, phase transitions are ubiquitous in living and cognitive systems. Di Paolo et al. characterize adaptivity in terms of phase transitions. An adaptive act is, they contend, a phase transition in which an agent undergoes a change in structure switching from an existing dynamical trajectory that would lead to a loss of viability eventually if left unchecked. The history of an organism can be described as the "cumulative change" in the configuration of the phase space that describes the behavior of the organism over the course of its lifetime.

Di Paolo et al. argues that this characteristic of path-dependence, whereby the agent's internal dynamics are dependent on its past history of phase transitions, is fundamentally incompatible with the idea of an attracting set of non-equilibrium steady-states to which the organism repeatedly returns when perturbed. A system that conserves its organization in this way will, they argue, quickly forget its history. The long-term average of the states the system visits over time will be equivalent to the averaging of the states in an ensemble of the system at a time. Di Paolo et al. take this to describe a key difference between physical systems that tend to conserve invariant structure and biological systems that rely upon a continuous reconfiguration of their structure following critical transitions. If adaptivity happens in such moments of

[11] Our argument that agents with sensorimotor autonomy will circumvent the problem of meaning shares much in common with the account of relevance realization developed by John Vervaeke et al. in a number of publications (e.g., Vervaeke et al., 2012; Vervaeke and Ferraro, 2013). Vervaeke et al. understand relevance realization in terms of the self-organizing optimisation of trade-offs between opponent yet complementary learning strategies. An example is how relevance could be realized in relation to the goal of threat avoidance through optimizing the trade-off between fight and flight. Vervaeke et al. also frequently use the example of the trade-off between exploration and exploitation. A discussion of similarities and differences between our approaches is unfortunately beyond the scope of this article. However, see Hovhannisyan and Vervaeke (2021) for a recent account of how the concept of relevance realization could contribute to developing an enactive approach to humanistic psychology.

[12] The ergodicity assumption is employed as an approximation to model systemic behavior. It requires that a system returns to approximately the same states over time. The notion of approximate similarity however should not be mistaken for numerical identity.

critical transition, it would seem to follow that adaptivity cannot be understood in terms of adaptive active inference.

First, let us agree with Di Paolo et al. that adaptivity does indeed occur in moments of critical transition in the dynamics of an organism-environment system (cf. Varela, 1995). Indeed historical path-dependence has been central to how we have analyzed meaning in this paper. We suggest the appearance of incompatibility of adaptivity, so conceived, with the FEP may stem from the generality of the FEP. Recall how the FEP is equally applicable both to physical systems that engage in mere active inference, and to biological systems which engage in adaptive active inference. The past history of dynamical interaction is indeed irrelevant to describing how the swinging pendulums enslave each other over time. However, the path independence of behavior is less obviously true of systems that exhibit adaptive active inference.

Recall that such systems are able to strike the right balance between the reduction of expected free energy through instrumental actions, and temporarily increasing free energy through exploration of the environment. To strike a balance between exploitation and exploration an adaptive active inference agent will need to instantiate a metastable dynamics. Metastability is the consequence of two competing tendencies (Kelso, 1995): the tendency of the parts of the system to separate and express their own intrinsic dynamics, which leads to an increase in free energy, and the tendency of the parts to integrate and coordinate to create new dynamics, in the way that Di Paolo et al. argue is required for adaptivity. Metastable systems are able to transit between regions of their phase space spontaneously without external perturbation. The structure of a metastable system is therefore transient. Systems with metastable dynamics avoid getting trapped in fixed-point attractors that lead to a single outcome. The internal dynamics are instead itinerant or wandering in a way that allows for exploratory behaviors that temporarily increase free energy (Zarghami and Friston, 2020). However, such temporary increases in free energy allow for just the kind of dynamical reconfiguration that Di Paolo et al. take to be essential for adaptivity.[13]

Indeed we suggest that systems that can find the right balance between reducing and temporarily inducing increases in entropy would need to be capable of dynamically reconfiguring their internal dynamics in ways that fit with the context in which they are acting. This is not to deny that the internal dynamics of an adaptive active inference agent can never become rigid and inflexible over time. However such rigidity is perhaps a signature feature of psychopathologies (cf. Carhart-Harris et al., 2014). Think for instance of obsessive compulsive disorder in which the agent finds themselves trapped in maladaptive cycles of behavior. What is characteristic of such pathological behaviors is a weakening of metastable dynamics that in healthy individuals allows for finding the right balance between reducing and increasing entropy.

To summarize our response to Di Paolo et al., we have argued that an agent that exhibits adaptive active inference will exhibit the historical path-dependence of behavior they take to be required for adaptivity. Such an agent will need to exhibit path-dependent behavior if it is to succeed in maximizing both the instrumental and epistemic value of its action policies. Indeed, any system that learns a model of its environment will exhibit plastic changes in its internal dynamics. The appearance of an incompatibility between the enactive approach to life and cognition and the FEP stems from the generality of the FEP. Certainly some of the systems to which the FEP applies will not be capable of adaptivity (e.g., those that are modeled as performing mere adaptive inference) but it doesn't follow that no systems the FEP is used to model could exhibit adaptivity.

We turn next to a second recent paper that also challenges our proposal to use adaptive active inference to formally model sensorimotor autonomy. It has been argued that to apply the mathematics of the FEP to concrete physical systems requires specific assumptions that do not typically apply to the sensorimotor interactions of living systems (Aguilera et al., 2021). Aguilera et al. argue for the opposite conclusion from the one we have been defending, that the FEP is highly particular in the systems to which it applies. Indeed they claim the FEP is so particular in its requirements as to fail to pick out the class of systems that would qualify as having sensorimotor autonomy. Aguilera et al. make their argument by considering the assumptions that would be required to apply the FEP to a class of simple systems whose dynamics are described by stochastic linear differential equations. They select such systems on the grounds that if the assumptions of the FEP do not apply to such simple systems, it is unlikely that they hold for more complex non-linear systems.

Aguilera et al. begin by considering the type of sensorimotor interface that, according to the FEP, mediates the interaction of the internal dynamics of the agent and the external dynamics of the environment.[14] They show that the sensorimotor interface must have two statistical properties. First, they must be described by the Markov blanket formalism, whereby internal and external states are conditionally independent given the sensory and active states of a Markov blanket. Second, the sensorimotor interface must be such that solenoidal couplings between internal and external states are decoupled by blanket states. Aguilera et al. define "solenoidal couplings" as arising from "dissipative tendencies in the system" that drive a system "away from equilibrium" (Aguilera et al., 2021: p.2). They show that any

[13] One of our reviewers objected that random dynamical systems that tend toward a NESS cannot possess metastable dynamics. Friston has however provided many models of active inference that in his words "provide a key connection to dynamical approaches to the brain that emphasize…metastability" (Friston, 2010, p.134), a connection that the reviewer takes to be excluded by the derivation of the FEP. We suspect the reviewer is confusing the mathematics that are used to derive the FEP, which may well contain equations that fail to capture metastability, with the use of the FEP to model systems with metastable dynamics. We will return to this point below in responding to the challenges raised by Aguilera et al. recent work.

[14] Aguilera et al. discuss a second assumption required for applying the FEP to concrete systems that "implies decoupling the actions of an agent from its history of previous states" (Aguilera et al., 2021, p.3). Their critique of this assumption is related to that of Di Paolo et al., discussed earlier in this section, but it also raises additional issues we cannot tackle in this paper but hope to return to in future work.

system that possesses a sensorimotor interface satisfying these two statistical properties will exhibit an internal dynamics that can be described in terms of descent on a free energy gradient. Aguilera et al. show that to connect the average flow or internal dynamics of a system with a gradient minimizing free energy requires the assumption that a Markov blanket precludes solenoidal couplings between internal and external dynamics.

The no solenoidal couplings (NSC) assumption raises difficulties for our claim that adaptive active inference is sufficient for adaptivity. Aguilera et al. show that a system that conforms with the NSC assumption will possess a sensorimotor interface that precludes adaptivity. This is because systems that satisfy the NSC assumption will possess a sensorimotor interface that permits only fully symmetrical interaction loops to form between agents and environments. If systems that conform with the FEP must exhibit fully symmetrical sensorimotor interactions with the environment, such systems will lack adaptivity. For adaptivity, as we have seen above, requires interactional asymmetry between agent and environment. Adaptivity requires that the agent be able to modulate its interaction with the environment in such a way as to influence the constraints on the agent's behavior, where some of these constraints are due to the agent, and others to its environment.[15]

By way of a reply, we begin by briefly considering more carefully the claim that the sensorimotor interface implied by the FEP can be modeled as a Markov blanket that induces a separation described statistically in terms of conditional independence of internal and external states. Now it is crucial to note that the Markov blanket is not fixed once and for all but the sensory and active states out of which it is built continuously undergo change, based on the agent's coupling with its environment. The accumulation of fluctuations will gradually render the states of the Markov blanket independent of the initial conditions that gave rise to them. Given sufficient time, the FEP implies that a system that minimizes expected free energy should instantiate a probability density that converges on a NESS. However in the intervening period of time as fluctuations accumulate, internal and external dynamics enter into a transient state of conditional dependence mediated by the Markov blanket. Thus, the Markov blanket condition, that is the conditional independence of internal and external dynamics, is temporarily violated. This violation of the Markov blanket condition has been argued to allow for memory (Parr et al., 2021) but we suggest it should also allow for a modulation of the agent-environment relation in line with interaction asymmetry.

This takes us back to our earlier discussion of the historical path dependence of behavior. Recall that it was the capacity of agents that conform with the FEP to modulate the parameters and constraints on their coupling with the environment that was in contention in this earlier discussion. We argued that neural processes that alter their dynamics in fluid and adaptive ways, in response to the requirements of particular contexts of activity, are part and parcel of adaptive active inference. Such neural processes are an essential part of selecting action policies that maximize instrumental and epistemic value in a dynamical environment. The model of chemotaxis of Tschantz et al. already exhibits a bistable dynamical profile. It is able to endogenously switch between running and tumbling based on changes in free energy. We take this simulation as a demonstration that an agent can be formally described in accordance with the FEP and exhibit a minimal form of sensorimotor agency.

Aguilera et al. may respond that our argument fails since systems that satisfy the NSC assumption *must* engage in symmetrical sensorimotor interactions with the environment. We suggest however that the systems that the FEP models are dynamical systems that can temporarily violate the assumptions the models rest upon, while at the same time on average and over time conforming to those assumptions. Aguilera and colleagues ask what assumptions are needed to apply the equations of FEP to a specific class of systems whose dynamics are described by stochastic linear differential equations. Such an argument seems to assume however that in order for the FEP to be used to represent the dynamics of physical systems, its mathematical equations must literally be instantiated by those physical systems. This is an example of what we have elsewhere called the "literalist fallacy"—the fallacy of taking the properties of FEP models to literally map onto real-world target systems (Kirchhoff et al., 2022). We suggest instead that active inference models based on the FEP are better conceived of as idealisations and approximations that introduce deliberate distortions. The Markov blanket assumption is an example of such a distortion, which is why the systems that the FEP describes can violate this assumption, while at the same time FEP based models can accurately represent the longer-term dynamics of such systems.

Similar arguments can be made in response to the argument of Di Paolo et al. that systems with an attracting set or NESS are memoryless, and are therefore incapable of historical path-dependent behavior. Di Paolo et al. critique trades on the assumption that in order for the FEP to truthfully represent a system, the properties it models must literally be instantiated by a system. We have been arguing however that the systems the FEP purports to model are dynamical systems that can fruitfully be represented as tending to evolve toward states belonging to their attracting set. The FEP can serve as the basis for models that provide truthful but approximate and idealized representations of such systems, including systems that instantiate sensorimotor autonomy, if the arguments of our paper are valid. We conclude with some additional issues for further research.

## CONCLUSION

Artificial intelligence from its earlier days has struggled with the problem of meaning. The information that computers process does not mean anything *for* the system that is doing the processing. This information only means something for the users of these systems. We have argued that the imperative to minimize expected free energy could serve as an intrinsic

[15]Di Paolo et al. note that the individual agent need not always be the source of the modulation of its coupling. Other agents can also induce asymmetric changes in dynamical constraints resulting in a modulation of the individual's coupling with the environment (Di Paolo et al., 2017: p.120). We set aside this important complication here.

norm for an artificial agent. Thus adaptive active inference could provide a formal description of the conditions an artificial agent would need to satisfy to possess sensorimotor autonomy and thus to perceive a meaningful environment (see Kolchinsky and Wolpert, 2018 for a related proposal). We finished up by considering two objections to our thesis that the imperative to minimize expected free energy may serve as an intrinsic norm for an agent. These objections generate a number of important questions for further research, which we will end by highlighting.

First, we have argued that models of adaptive active inference can be used to formally describe systems that possess sensorimotor autonomy. However it could be objected that such an agent could indeed be considered a *model* of sensorimotor autonomy but without itself possessing this property. Just as a model of intelligence may lack intelligence, similarly a model of sensorimotor autonomy may not itself instantiate this property. To genuinely instantiate such a property, it might be argued that an agent would need to have a material body composed of processes that self-organize to form operationally closed networks, and that distinguish the agent as a unified individual from its environment. The artificial agent of Tschantz et al., which we have taken as our main example in this paper, has no material body but exists only *in silico.* When it is simulating chemotaxis, it does not engage in exchanges of matter and energy with its environment that are part of its process of self production and self differentiation. Thus no matter how good a model of autonomy and adaptivity it may be, it might be argued it does not yet possess these organizational properties.

Second, and relatedly, Froese and Taguchi (2019) have argued that modeling autonomy and adaptivity will fail to solve the problem of meaning. They concede that artificial agents may be simulated that act *as if* they have their own intrinsic norms. They argue however that an important disanalogy will remain with organic life. An organism actively brings about its own existence through engaging in metabolic activity. Its continued existence or being is, in an important sense, a consequence of its own doings. It is this relationship between being and doing that makes for goals and concerns that are intrinsic to the organism. Froese and Taguchi (2019) argue that any simulation of artificial agency cannot be said to genuinely have a existence that is the consequence of its own doing. They argue that there is no room for meaning, normativity or value to make a difference to the behavior of such agents insofar as they act in a simulated environment that is fully deterministic. The behavior of a simulated agent is due to dynamical constraints on its internal and interactional dynamics, not to the agent's bringing forth a domain of meaningful action. Froese and Taguchi argue on this basis that if meaning is to make a real difference to the behavior of an agent, some indeterminacy must be built into the agent's engagement with its environment.

Finally, more work is needed on the challenges that arise from applying the mathematics of the FEP to concrete sensorimotor agents. Are systems whose dynamics are describable in terms of non-equilibrium steady-states also capable of path-dependent behaviors, as we have argued? If the application of the FEP to concrete systems depends upon the NRC assumption, as Aguilera et al. show, does it follow that all systems describable in terms of the FEP must engage in symmetrical interactions with their environment? Can the FEP be used to model systems with metastable dynamics? We argued that these are related challenges but more work is certainly required on the implications of answering them for the FEP. While there is a good deal more work to be done, we have argued that the synthesis of enactive ideas with the FEP may set biologically inspired AI research on a promising path for addressing the problem of meaning.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

JK wrote the manuscript. MK and TF provided critical feedback and editing suggestions. All authors contributed to the article and approved the submitted version.

## FUNDING

JK was supported by grants awarded to Erik Rietveld from the Dutch Organisation for Scientific Research and the European Research Council H2020 Starting Grant (number 679190). MK was supported by the Australian Research Council Discovery Project (DP170102987).

## REFERENCES

Aguilera, M., Millidge, B., Tschantz, A., and Buckley, C. L. (2021). How particular is the physics of the free energy principle? *Phys. Life Rev.* 40, 24–50. doi: 10.1016/j.plrev.2021.11.001

Allen, M., and Tsakiris, M. (2018). "The body as first prior: interoceptive predictive processing and the primacy of self-models," in *The Interoceptive Mind: From Homeostasis to Awareness*, eds M. Tsakiris and H. De Preester (Oxford, UK: Oxford University Press), 27–45. doi: 10.1093/oso/9780198811930.003.0002

Anderson, M. L. (2014). *After Phrenology: Neural Reuse and The Interactive Brain*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/10111.001.0001

Anderson, M. L., Richardson, M. J., and Chemero, A. (2012). Eroding the boundaries of cognition: implications of embodiment. *Top. Cogn. Sci.* 4, 717–730. doi: 10.1111/j.1756-8765.2012.01211.x

Andrews, M. (2021). The math is not the territory: navigating the Free Energy Principle. *Biol. Philos.* 36, 30. doi: 10.1007/s10539-021-09807-0

Baltieri, M., Buckley, C. L., and Bruineberg, J. (2020). "Predictions in the eye of the beholder: an active inference account of Watt governors," in *Artificial Life Conference Proceedings*, eds J. Bongard, J. Lovato, L. Herbert-Dufresne, R, Dasari, and L. Soros (Cambridge, MA: MIT Press), 121–129. doi: 10.1162/isal_a_00288

Barandiaran, X. E. (2008). *Mental Life: A Naturalised Approach to the Autonomy of Cognitive Agents* (Ph.D. thesis). UPV-EHU, University of the Basque Country.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry and spatio-temporality in action. *Adapt. Behav.* 17, 367–386. doi: 10.1177/1059712309343819

Bruineberg, J., Dolega, K., Dewhurst, J., and Baltieri, M. (2022). The emperor's new Markov blankets. *Behav. Brain Sci.* 1–63. doi: 10.1017/S0140525X21002351

Bruineberg, J., Kiverstein, J., and Rietveld, E. (2018). The anticipating brain is not a scientist: the free energy principle from an ecological-enactive perspective. *Synthese* 195, 2417–2444. doi: 10.1007/s11229-016-1239-1

Carhart-Harris, R., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., et al. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front. Hum. Neurosci.* 8, 20. doi: 10.3389/fnhum.2014.00020

Colombo, M., and Wright, C. (2018). First principles in the life-sciences: the free-energy principle, organicism and mechanism. *Synthese* 198, 3463–3488. doi: 10.1007/s11229-018-01932-w

Da Costa, L., Lanillos, P., Sajid, N., Friston, K., and Khan, S. (2022). How active inference could help revolutionise robotics. *Entropy* 24, 361. doi: 10.3390/e24030361

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9540.001.0001

Dennett, D. (1984). "Cognitive wheels: the frame problem of AI," in *Minds, Machines and Evolution,* ed C. Hookway (Cambridge, MA: Cambridge University Press), 129–151.

Di Paolo, E., Thompson, E., and Beer, R. (2022). Laying down a forking path: incompatibilities between enaction and the free energy principle. *Philos. Mind Sci.* 3, 1–39. doi: 10.33735/phimisci.2022.9187

Di Paolo, E. A. (2003). "Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop," in *Dynamic Systems Approach for Embodiment and Sociality: From Ecological Psychology to Robotics*, eds K. Murase and T. Asakura (Adelaide, SA: Advanced Knowledge International), 19–42.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenol. Cogn. Sci.* 4, 429–452. doi: 10.1007/s11097-005-9002-y

Di Paolo, E. A. (2018). "The enactive conception of life," in *The Oxford Handbook of 4E (Embodied, Embedded, Enactive, Extended) Cognition,* eds A. Newen, L. De Bruin and S. Gallagher. (Oxford, UK: Oxford University Press), 71–95. doi: 10.1093/oxfordhb/9780198735410.013.4

Di Paolo, E. A., Buhrmann, T., and Barandiaran, X. E. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford, UK: Oxford University Press. doi: 10.1093/acprof:oso/9780198786849.001.0001

Di Paolo, E. A., and Thompson, E. (2014). "The enactive approach," in *The Routledge Handbook of Embodied Cognition,* ed L. Shapiro (London: Routledge), 68–78.

Dreyfus, H. L. (1992). *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Egbert, M., and Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Front. Hum. Neurosci.* 8, 590. doi: 10.3389/fnhum.2014.00590

Fodor, J. (2000). *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/4627.001.0001

Fotopoulou, A., and Tsakiris, M. (2017). Mentalising homeostasis: the social origins of interoceptive inference. *Neuropsychoanalysis* 19, 3–28. doi: 10.1080/15294145.2017.1294031

Friston, K. (2011). "Embodied inference: Or I think therefore I am, if I am what I think," in *The implications of embodiment (Cognition and Communication)*, eds W. Tschacher, and C. Bergomi (Exeter, UK: Imprint Academic), 89–125.

Friston, K. (2019). A free energy principle for a particular physics. *arXiv*, 1906.10184. Available online at: https://arxiv.org/abs/1906.10184.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K. J. (2013). Life as we know it. *J. R. Soc. Interface* 10, 20130475. doi: 10.1098/rsif.2013.0475

Friston, K. J., and Ao, P. (2012). Free energy, value and attractors. *Comput. Math. Models Med.* 2012, 937860, doi: 10.1155/2012/937860

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S., (2017b). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Froese, T., and Taguchi, S. (2019). The problem of meaning in AI and Robotics: still with us after all these years. *Philosophies* 4, 14. doi: 10.3390/philosophies4020014

Froese, T., and Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artif. Intell.* 173, 366–500. doi: 10.1016/j.artint.2008.12.001

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K., and Ramstead, M. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Comput*. 33, 398–446. doi: 10.1162/neco_a_01341

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199682737.001.0001

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Hovhannisyan, G., and Vervaeke, J. (2021). Enactivist big five theory. *Phenomenol. Cogn. Sci.* 21, 341–375. doi: 10.1007/s11097-021-09768-5

Hutto, D., and Myin, E. (2013). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262018548.001.0001

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free energy principle. *PLOS: Comput. Biol.* 9, e1003094. doi: 10.1371/journal.pcbi.1003094

Kauffman, S. (2019). *A World Beyond Physics: The Emergence and Evolution of Life.* Oxford: Oxford University Press.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behaviour*. Cambridge, MA: MIT Press.

Kirchhoff, M., Kiverstein, J., and Robertson, I., (2022). The literalist fallacy and the free energy principle: On model-building, scientific realism and instrumentalism. *Br. J. Philos. Sci.*

Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese* 195, 2519–2540. doi: 10.1007/s11229-016-1100-6

Kirchhoff, M. D., and Froese, T. (2017). Where there is life there is mind: In support of the strong life-mind continuity thesis. *Entropy* 19, 169. doi: 10.3390/e19040169

Kirchhoff, M. D., and Kiverstein, J. (2019). How to determine the boundaries of the mind: a markov blanket proposal. *Synthese* 1–20. doi: 10.1007/s11229-019-02370-y

Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K. J., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15, 20170792. doi: 10.1098/rsif.2017.0792

Kiverstein, J., and Kirchhoff, M. D. (2022). Scientific realism about Friston blankets without literalism: Commentary on Bruineberg et al. The Emperor's New Markov Blankets. *Behav. Brain Sci.* (Forthcoming)

Kiverstein, J., Miller, M., and Rietveld, E., (2019). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese* 196, 2847–2869. doi: 10.1007/s11229-017-1583-9

Kiverstein, J., and Rietveld, E. (2018). Reconceiving representation-hungry cognition: an ecological-enactive perspective. *Adapt. Behav.* 26, 147–163. doi: 10.1177/1059712318772778

Kolchinsky, A., and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* 8, 20180041. doi: 10.1098/rsfs.2018.0041

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40, 1–72. doi: 10.1017/S0140525X16001837

Lanollis, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., et al. (2021). *Active inference in robotics and artificial intelligence: survey and challenges*. Available online at: https://arxiv.org/abs/2112.01871 (accessed May, 01 2022).

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realisation of the Living* (Dordrecht). The Netherlands: D. Reidel Publishing.

Millidge, B., Tschantz, A., and Buckley, C. L. (2021). Whence the expected free energy? *Neural Comput.* 33, 447–482. doi: 10.1162/neco_a_01354

Nave, K. (2022). Boundaries and borders gone! But life goes on. Commentary on Bruineberg et al. The emperor's new Markov blankets, forthcoming in Behavioural and Brain Sciences. Preprint. Available online at: https://psyarxiv.com/h3zpx (accessed February 3, 2022).

Oyama, S. (2000). *The Ontogeny of Information: Developmental Systems and Evolution, 2nd Edn*, revised and expanded, Durham, NC: Duke University Press. doi: 10.2307/j.ctv1220mm5

Parr, T., Da Costa, L., Heins, C., Ramstead, M. J. D., and Friston, K. J. (2021). Memory and markov blankets. *Entropy* 23, 1105. doi: 10.3390/e23091105

Parr, T., and Friston, K. (2017). Uncertainty, epistemics and active inference. *J. Ro. Soc. Interface* 14, 20170376. doi: 10.1098/rsif.2017.0376

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann. doi: 10.1016/B978-0-08-051489-5.50008-4

Pezzulo, G., Rigoli, F., and Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* 22, 294–306. doi: 10.1016/j.tics.2018.01.009

Raja, V., Valluri, D., Baggs, E., Chemero, A., and Anderson, M. L. (2021). The Markov blanket trick: on the scope of the free energy principle and active inference. *Phys. Life Rev.* 39. 49–72. doi: 10.1016/j.plrev.2021.09.001

Ramírez-Vizcaya, S., and Froese, T. (2020). "Agents of habit: refining the artificial life route to artificial intelligence," in *Proceedings of the ALIFE 2020: The 2020 Conference on Artificial Life* (ASME), 78–86. doi: 10.1162/isal_a_00298

Ramstead, M., Friston, K., and Hipólito, I., (2020b). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy* 22, 889. doi: 10.3390/e22080889

Ramstead, M., Kirchhoff, M. D., Constant, A., and Friston, K. J. (2021). Multiscale integration: beyond internalism and externalism. *Synthese* 198, 41–70. doi: 10.1007/s11229-019-02115-x

Ramstead, M., Kirchhoff, M. D., and Friston, K. J. (2020a). A tale of two densities: active inference is enactive inference. Adapt. Behav. 28, 225–239. doi: 10.1177/1059712319862774

Samuels, R. (2010). Classical computationalism and the many problems of cognitive relevance. *Stud. History Philos. Sci. Part A* 41, 280–293. doi: 10.1016/j.shpsa.2010.07.006

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Seth, A., and Tsakiris, M. (2018). Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* 22, 969–981. doi: 10.1016/j.tics.2018.08.008

Shanahan, M. (2016). "The frame problem," in *The Stanford Encyclopaedia of Philosophy*. Available online at: https://plato.stanford.edu/entries/frame-problem/

Suzuki, K., Miyahara, K., and Miyazono, K. (2022). Who tailors the blankets? Commentary on Bruineberg et al. The emperor's new Markov blankets, forthcoming in Behavioural and Brain Sciences. Preprint, Available online at: https://psyarxiv.com/2sryv/ (accessed February 3, 2022).

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Tschantz, A., Seth, A. K., and Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Comput. Biol.* 16, e1007805. doi: 10.1371/journal.pcbi.1007805

Van de Cruys, S. (2017). "Affective value in the predictive mind," in *Philosophy of Predictive Processing*, eds T. Metzinger and W. Wiese (Frankfurt am Main: OPEN MIND), 1–21.

van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adapt. Behav.* 29, 315–329 doi: 10.1177/1059712320918678

van Es, T., and Hipólito, I. (2020). Free energy principle, computationalism and realism: a tragedy. Available online at: http://philsci-archive.pitt.edu/18497/ (accessed February 3, 2022).

van Es, T., and Kirchhoff, M. D. (2021). Between pebbles and organisms: weaving autonomy into the Markov blanket. *Synthese* 199, 6623–6644. doi: 10.1007/s11229-021-03084-w

Varela, F.J. (1991). "Organism: a meshwork of selfless selves," in *Organism and the Origins of the Self,* ed A. I. Tauber (Dordrecht: Kluwer), 79–107. doi: 10.1007/978-94-011-3406-4_5

Varela, F. J. (1979). *Principles of Biological Autonomy*. New York, NY: North Holland

Varela, F. J. (1995). "The re-enchantment of the concrete: Some biological ingredients for a nouvelle cognitive science," in *The Artificial Life Route to Artificial Intelligence*, eds L. Steels and R. Brooks (Hove, UK: Lawrence Erlbaum Associates), 11–22.

Vervaeke, J., and Ferraro, L. (2013). "Relevance realization and the neurodynamics and neuroconnectivity of general intelligence," in *Smart Data,* eds I. Harvey, A. Cavoukian, G. Tomko, D. Borrett, H. Kwan, and D. Hatzinakos (New York, NY: Springer), 57–68. doi: 10.1007/978-1-4614-6409-9_6

Vervaeke, J., Lillicrap, T. P., and Richards, B. A. (2012). Relevance realization and the emerging framework in cognitive science. *J. Logic Comput.* 22, 79–99. doi: 10.1093/logcom/exp067

Von Uexküll, J. (1957). "A stroll through the worlds of animals and men," in *Instinctive Behaviour: The Development of a Modern Concept,* ed K. S. Lashley (New York, NY: International Universities Press), 5–80.

Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/5824.001.0001

Zarghami, T. S., and Friston, K. J. (2020). Dynamic effective connectivity. *Neuroimage* 207, 116453. doi: 10.1016/j.neuroimage.2019.116453

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Kiverstein, Kirchhoff and Froese. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers | Frontiers in Neurorobotics

TYPE Hypothesis and Theory
PUBLISHED 28 July 2022
DOI 10.3389/fnbot.2022.896229

Check for updates

OPEN ACCESS

EDITED BY
Adam Safron,
Johns Hopkins Medicine, United States

REVIEWED BY
Valerio Santangelo,
University of Perugia, Italy
Nicholas Huang,
Biofourmis Inc., United States

*CORRESPONDENCE
Ajith Anil Meera
a.anilmeera@tudelft.nl
Filip Novicky
filip.novicky@donders.ru.nl

†These authors have contributed equally to this work and share first authorship
‡These authors have contributed equally to this work

RECEIVED 14 March 2022
ACCEPTED 28 June 2022
PUBLISHED 28 July 2022

CITATION
Anil Meera A, Novicky F, Parr T, Friston K, Lanillos P and Sajid N (2022) Reclaiming saliency: Rhythmic precision-modulated action and perception.
*Front. Neurorobot.* 16:896229.
doi: 10.3389/fnbot.2022.896229

COPYRIGHT
© 2022 Anil Meera, Novicky, Parr, Friston, Lanillos and Sajid. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Reclaiming saliency: Rhythmic precision-modulated action and perception

Ajith Anil Meera[1*†], Filip Novicky[2*†], Thomas Parr[3], Karl Friston[3], Pablo Lanillos[4‡] and Noor Sajid[3‡]

[1]Department of Cognitive Robotics, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, Netherlands, [2]Department of Neurophysiology, Donders Institute for Brain Cognition and Behavior, Radboud University, Nijmegen, Netherlands, [3]Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom, [4]Department of Artificial Intelligence, Donders Institute for Brain Cognition and Behavior, Radboud University, Nijmegen, Netherlands

Computational models of visual attention in artificial intelligence and robotics have been inspired by the concept of a saliency map. These models account for the mutual information between the (current) visual information and its estimated causes. However, they fail to consider the circular causality between perception and action. In other words, they do not consider where to sample next, given current beliefs. Here, we reclaim salience as an active inference process that relies on two basic principles: uncertainty minimization and rhythmic scheduling. For this, we make a distinction between attention and salience. Briefly, we associate attention with precision control, i.e., the confidence with which beliefs can be updated given sampled sensory data, and salience with uncertainty minimization that underwrites the selection of future sensory data. Using this, we propose a new account of attention based on rhythmic precision-modulation and discuss its potential in robotics, providing numerical experiments that showcase its advantages for state and noise estimation, system identification and action selection for informative path planning.

KEYWORDS

**attention, saliency, free-energy principle, active inference, precision, brain-inspired robotics, cognitive robotics**

## 1. Introduction

Attention is a fundamental cognitive ability that determines which events from the environment, and the body, are preferentially processed (Itti and Koch, 2001). For example, the motor system directs the visual sensory stream by orienting the fovea centralis (i.e., the retinal region of highest visual acuity) toward points of interest within the visual scene. Thus, the confidence with which the causes of sampled visual information are inferred is constrained by the physical structure of the eye—and eye movements are necessary to minimize uncertainty about visual percepts (Ahnelt, 1998). In neuroscience, this can be attributed to two distinct, but highly interdependent attentional processes: (*i*) attentional gain mechanisms reliant on estimating the sensory precision of current data (Feldman and Friston, 2010; Yang et al., 2016a), and

(*ii*) attentional salience that involves actively engaging with the sensorium to sample appropriate future data (Lengyel et al., 2016; Parr and Friston, 2019). Here we refer to perceptual-related salience, i.e., processing of low-level visual information (Santangelo, 2015). Put simply, we formalize the fundamental difference between attention—as optimizing perceptual processing—and salience as optimizing the sampling of what is processed. This highlights the dynamic, circular nature with which biological agents acquire, and process, sensory information.

Understanding the computational mechanisms that undergird these two attentional phenomena is pertinent for deploying apt models of (visual) perception in artificial agents (Klink et al., 2014; Mousavi et al., 2016; Atrey et al., 2019) and robots (Frintrop and Jensfelt, 2008; Begum and Karray, 2010; Ferreira and Dias, 2014; Lanillos et al., 2015a). Previous computational models of visual attention, used in artificial intelligence and robotics, have been inspired (and limited) by the feature integration theory proposed by Treisman and Gelade (1980) and the concept of a saliency map (Tsotsos et al., 1995; Itti and Koch, 2001; Borji and Itti, 2012). Briefly, a saliency map is a static two-dimensional 'image' that encodes stimulus relevance, e.g., the importance of particular region. These maps are then used to isolate relevant information for control (e.g., to direct foveation of the maximum valued region). Accordingly, computational models reliant on this formulation do not consider the circular-dependence between action selection and cue relevance—and simply use these static saliency maps to guide action.

In this article, we adopt a first principles account to disambiguate the computational mechanisms that underpin attention and salience (Parr and Friston, 2019) and provide a new account of attention. Specifically, our formulation can be effectively implemented for robotic systems and facilitates both state-estimation and action selection. For this, we associate attention with precision control, i.e., the confidence with which beliefs can be updated given (current) sampled sensory data. Salience is associated with uncertainty minimization that influences the selection of future sensory data. This formulation speaks to a computational distinction between action selection (i.e., where to look next) and visual sampling (i.e., what information is being processed). Importantly, recent evidence demonstrates the rhythmic nature of these processes *via* a theta-cycle coupling that fluctuates between high and low precision—as unpacked in Section 2. From a robotics perspective, resolving uncertainty about states of affair speaks to a form of Bayesian optimality, in which decisions are made to maximize expected information gain (Lindley, 1956; Friston et al., 2021; Sajid et al., 2021a). The duality between attention and salience is important for resolving uncertainty and enabling active perception. Significantly, it addresses an important challenge for defining autonomous robotics systems that can balance optimally between data assimilation (i.e., confidently perceiving current observations) and exploratory behavior to maximize information gain (Bajcsy et al., 2018).

In what follows, we review the neuroscience of attention and salience (Section 2) to develop a novel (computational) account of attention based on precision-modulation that underwrites perception and action (Section 3). Next, we face-validate our formulation within a robotics context using numerical experiments (Section 4). The robotics implementation instantiates a free energy principle (FEP) approach to information processing (Friston, 2010). This allows us to modulate the (appropriate) precision parameters to solve relevant robotics challenges in perception and control; namely, state-estimation (Section 4.2.2), system identification (Section 4.2.3), planning (Section 4.3), and active perception (Section 4.3.3). We conclude with a discussion of the requisite steps for instantiating a full-fledged computational model of precision-modulated attention—and its implications in a robotics setting.

## 2. Attention and salience in neuroscience

Our interactions with the world are guided by efficient gathering and processing of sensory information. The quality of these acquired sensory data is reflected in attentional resources that select sensations which influence our beliefs about the (current and future) states of affairs (Lengyel et al., 2016; Yang et al., 2016b). This selection is often related to gain control, i.e., an increase of neural spikes when an object is attended to. However, gain control only accounts for half the story because we can only attend to those objects that are within our visual field. Accordingly, if a salient object is outside the center of our visual field, we orient the fovea to points of interest. This involves two separate, but often conflated, processes: attention and salience—where the former relates to processing current visual data, and the latter to ensuring the agent samples salient data in the future (Parr and Friston, 2019). That these two processes are strongly coupled is exemplified by the pre-motor theory of attention (Rizzolatti et al., 1987), which highlights the close relationship between overt saccadic sampling of the visual field and the covert deployment of attention in the absence of eye movements. Specifically, it posits that covert attention[1] is realized *via* processes that are generated by particular eye movements but inhibits the action itself. In this sense, it does not distinguish between covert and overt[2] types of attention.

From a first principles (Bayesian) account, it is necessary to separate between attention and salience because they speak to different optimization processes. Explicitly, attention

1 Covert attention is where saccadic eye movements do not occur.

2 Overt attention deals with how an agent tracks the object with eye movements.

as a precision-dependent (neural) gain control mechanism that facilitates optimization of the *current* sampled sensory data (Desimone, 1996; Feldman and Friston, 2010). Conversely, salience is associated with selection of *future* data that reduces uncertainty (Friston et al., 2015; Mirza et al., 2016; Parr and Friston, 2019). Put simply, it is possible to optimize attention in the absence of eye movements and active vision, whereas salience is necessary to optimize the deployment of eye movements. In what follows, we formalize this distinction with a particular focus on visual attention (Kanwisher and Wojciulik, 2000), and discuss recent findings that speak to a rhythmic coupling that underwrites periodic deployment of gain control and saccades, *via* modulation of distinct precision parameters.

## 2.1. Attention as neural gain control

Neural gain control can be regarded as an amplifier of neural communication during attention tasks (Reynolds et al., 2000; Eldar et al., 2013). Computationally, this is analogous to modulating a precision term, or the inverse temperature parameter (Feldman and Friston, 2010; Parr and Friston, 2017a). For this reason, we refer to precision and gain control interchangeably. An increase in gain amplifies the postsynaptic responses of neurons to their pre-synaptic input. Thus, gain control rests on synaptic modulation that can emphasize—or preferentially select—a particular type of sensory data. From a Bayesian perspective (Rao, 2005; Spratling, 2008; Parr et al., 2018), this speaks to the confidence with which beliefs can be updated given sampled sensory data (i.e., optimal state estimation)—under a generative model (Whiteley and Sahani, 2008; Parr et al., 2018). For example, affording high precision to certain sensory inputs would lead to confident Bayesian belief updating. However, low precision reduces the influence of sensory input by attenuating the precision of the likelihood, relative to a prior belief, and current observations would do little to resolve ensuing uncertainty. Thus, sampled visual data (from different areas) can be predicted with varying levels of precision, where attention accentuates sensory precision. The deployment of precision or attention is influenced by competition between stimuli (i.e., which sensory data to sample) and prior beliefs. Interestingly, casting attention as precision or, equivalently, synaptic gain offers a coherency between biased competition (Desimone, 1996), predictive coding (Spratling, 2008) and generic active inference schemes (Feldman and Friston, 2010; Brown et al., 2013; Kanai et al., 2015; Parr et al., 2018).

Naturally, gain control is accompanied by neuronal variability, i.e., sharpened neural responses for the same task over time. Consistent with gain control, these fluctuations in neural responses across trials can be explained by precision engineered message passing (Clark, 2013) *via* (*i*) normalization models (Reynolds and Heeger, 2009; Ruff and Cohen, 2016), (*ii*) temperature parameter manipulation (Feldman and Friston, 2010; Parr and Friston, 2017a; Parr et al., 2018, 2019; Mirza et al., 2019), or (*iii*) introduction of (conjugate hyper-)priors that are either pre-specified (Sajid et al., 2020, 2021b) or optimized using uninformed priors (Friston et al., 2003; Anil Meera and Wisse, 2021). Recently, these approaches have been used to simulate attention by accentuating predictions about a given visual stimulus (Reynolds and Heeger, 2009; Feldman and Friston, 2010; Ruff and Cohen, 2016). For example, normalization models propose that every neuronal response is normalized within its neuronal ensemble (i.e., the surrounding neuronal responses) (Heeger, 1992; Louie and Glimcher, 2019). Thus, to amplify the neuronal response of particular neuron, the neuronal pool has to be inhibited such that particular neuron has a sharper evoked response (Schmitz and Duncan, 2018). Importantly, these (superficially distinct) formulations simulate similar functions using different procedures to accentuate responses over a particular neuronal pool for a given neuron or a group of neurons. This introduces shifts in precision to produce attentional gain and the precision of neuronal encoding.

## 2.2. Salience as uncertainty minimization

In the neurosciences, (visual) salience refers to the 'significance' of particular objects in the environment. Salience often implicates the superior colliculus, a region that encodes eye movements (White et al., 2017). This makes intuitive sense, as the superior colliculus plays a role in generation of eye movements—being an integral part of the brainstem oculomotor network (Raybourn and Keller, 1977)—and salient objects provide information that is best resolved in the center of the visual field, thus motivating eye movements to that location. For this reason, our understanding of salience is a quintessentially action-driving phenomenon (Parr and Friston, 2019). Mathematically, salience has been defined as Bayesian surprise (Itti and Koch, 2001; Itti and Baldi, 2009), intrinsic motivation (Oudeyer and Kaplan, 2009), and subsequently, epistemic value under active inference (Mirza et al., 2016; Parr et al., 2018). Active inference—a Bayesian account of perception and action (Friston et al., 2017a; Da Costa et al., 2020)—stipulates that action selection is determined by uncertainty minimization. Formally, uncertainty minimization speaks to minimization of an expected free energy functional over future trajectories (Da Costa et al., 2020; Sajid et al., 2021a). This action selection objective can be decomposed into epistemic and extrinsic value, where the former pertains to exploratory drives that encourage resolution of uncertainty by sampling salient observations, e.g., only checking one's watch when one does not know the time. However, after checking the watch there is little epistemic value in looking at it again. Generally, the tendency to seek out new locations—once uncertainty has been resolved at

the current fixation point—is called inhibition of return (Klein, 2000).

From an active inference perspective, this phenomenon is prevalent because a recent action has already resolved the uncertainty about the time and checking again would offer nothing more in terms of information gain (Parr and Friston, 2019). Accordingly, salience involves seeking sensory data that have a predictable, uncertainty reducing, effect on current beliefs about states of affairs in the world (Mirza et al., 2016; Parr et al., 2018). Thus salience contends with beliefs about data that must be acquired and the precision of beliefs about policies (i.e., action trajectories) that dictate it. Formally, this emerges from the imperative to maximize the amount of information gained regarding beliefs, from observing the environment. Happily, prior studies have made the connection between eye movements, salience, and precision manipulation (Friston et al., 2011; Brown et al., 2013; Crevecoeur and Kording, 2017). This connection emerges from planning strategies that allow the agent to minimize uncertainty by garnering the right kind of data.

Next, we consider recent findings on how the coupling of these two mechanisms, attention and salience, may be realized in the brain.

## 2.3. Rhythmic coupling of attention and salience

To illustrate the coupling between attention and salience, we turn to a recent rhythmic theory of attention. The theory proposes that coupling of saccades, during sampling of visual information, happens at neuronal and behavioral theta oscillations; a frequency of 3–8 Hz (Fiebelkorn and Kastner, 2019, 2021). This frequency simultaneously allows for: (*i*) a systematic integration of visual samples with action, and (*ii*) a temporal schedule to disengage and search the environment for more relevant information.

Given that gain control is related to increased sensory precision, we can accordingly relate saccadic eye movements to the decreased precision. This introduces saccadic suppression, a phenomenon that decreases visual gain during eye movements (Crevecoeur and Kording, 2017). This phenomenon was described by Helmholtz who observed that externally initiated eye movements (e.g., when oneself gently presses a side of an eye) eludes the saccadic suppression that accompanies normal eye movements—and we see the world shift, because optic flow is not attenuated (Helmholtz, 1925). An interesting consequence of this is that, as eye movements happen periodically (Rucci et al., 2018; Benedetto et al., 2020), there must be a periodic switch between high and low sensory precision, with high precision (or enhanced gain) during fixations and low precision (or suppressed gain) during saccades. Interestingly, it has been shown that rather than having action resetting the neural periodicity, it is better understood as something that aligns within an already existing rhythm (Hogendoorn, 2016; Tomassini et al., 2017). Additionally, the rhythmicity of higher and lower fidelity of sensory sampling has been shown to fluctuate rhythmically around 3 Hz (Benedetto and Morrone, 2017), suggesting that action emerges rhythmically when visual precision is low (Hogendoorn, 2016), triggering salience.

Building upon this, we hypothesize that theta rhythms generated in the fronto-parietal network (Fiebelkorn et al., 2018; Helfrich et al., 2018; Fiebelkorn and Kastner, 2020) couples saccades with saccadic suppression causing the switches between visual sampling and saccadic shifting. This introduces a diachronic aspect to the belief updating process (Friston et al., 2020; Parr and Pezzulo, 2021; Sajid et al., 2022); i.e., sequential fluctuations between attending to current data (perception) and seeking new data (action). This supports empirical findings that both eye movements (Sommer and Wurtz, 2006) and filtering irrelevant information (Phillips et al., 2016; Nakajima et al., 2019; Fiebelkorn and Kastner, 2020) are initiated in this cortical network. Interestingly, both eye movements and visual filtering then propagate to sub-cortical regions, i.e., the superior colliculus—for saliency map composition (White et al., 2017)—and the thalamus—for gain control (Kanai et al., 2015; Fiebelkorn et al., 2019), respectively. Furthermore, this is consistent with recent findings that the periodicity of neural responses are important for understanding the relation of motor responses and sensory information—i.e., perception-action coupling (Benedetto et al., 2020). Importantly, theta rhythms also speak to the speed (i.e., the temporal schedule) with which visual information is sampled from the environment (Busch and VanRullen, 2010; Dugué et al., 2015, 2016; Helfrich et al., 2018). Meaning visual information is not sampled continuously, as our visual experiences would suggest, but rather it is made of successive discrete samples (VanRullen, 2016; Parr et al., 2021).

The prefrontal theta rhythm has been associated with working memory (WM), a process that holds compressed information about the previously observed stimuli, in the sense that measured power in this frequency range using electroencephalography increases during tasks that place demands on WM (Axmacher et al., 2010; Hsieh and Ranganath, 2014; Köster et al., 2018; Brzezicka et al., 2019; Peters et al., 2020; Balestrieri et al., 2021; Pomper and Ansorge, 2021). The implication is that the neural processes that underwrite WM may depend upon temporal cycles with periods similar to that of perceptual sampling. Importantly, this cognitive process is influenced by how salient a particular stimulus was (Fine and Minnery, 2009; Santangelo and Macaluso, 2013; Santangelo et al., 2015). Moreover, WM has been implicated with attentional mechanisms (Knudsen, 2007; Gazzaley and Nobre, 2012; Oberauer, 2019; Peters et al., 2020; Panichello and Buschman, 2021). This is aligned with our account

where we illustrate a rhythmic coupling between salience and attention.

In summary, the computations that underwrite attention and active vision are coupled and exhibit circular causality. Briefly, selective attention and sensory attenuation optimize the processing of sensory samples and which particular visual percepts are inferred. In turn, this determines appropriateness of future eye movements (or actions) and shapes which prior stimuli are encoded into the agent's working memory. Interestingly, the close functional (and computational) link between the two mechanisms endorses the pre-motor theory of attention.

## 3. Proposed precision-modulated account of attention and salience

Here, we introduce our precision-modulated account of perception and action. A graphical illustration is provided in Figure 1. For this, we turn to attention and salient action selection which have their roots in biological processes relevant for acquiring task-relevant information. Under an active inference account, this attention influences (posterior) state estimation and can be associated with increased precision of belief updating and gain control—described in Section 2.1. Furthermore, this is distinct from salience despite interdependent neuronal composition and computations.

Further alignment between the two constructs can be revealed by considering the temporal scheduling between movement (i.e., action) and perception for uncertainty resolution (Parr and Friston, 2019). We postulate that this perception-action coupling is best understood as a periodic fluctuation between minimizing uncertainty and precision control. Subsequently, action is deployed to reduce uncertainty. Such an alignment specifies what stimulus is selected and under what level of precision it is processed. Parr and Friston (2019) hypothesize that action alignment with precision is due to the eye structure that provides precise information in the fovea and requires the agent to foveate the most informative stimulus. We extend this by considering the periodic deployment of gain control with saccades (Hogendoorn, 2016; Benedetto and Morrone, 2017; Tomassini et al., 2017; Fiebelkorn and Kastner, 2019; Nakayama and Motoyoshi, 2019).

Accordingly, our formulation defines attention as precision control and salience as uncertainty minimization supported by discrete sampling of visual information at a theta rhythm. This synchronizes perception and action together in an oscillatory fashion (Hogendoorn, 2016). Importantly, a Bayesian formulation of this can be realized as precision manipulation over particular model parameters. We reserve further details for Section 4.

**Summary** Based upon our review, we propose a precision-modulated account of attention and salience, emphasizing the diachronic realization of action and perception. In the following sections, we investigate a realization of this model for a robotic system.

## 4. Precision-based attention for Robotics

The previous section introduced a conceptual account to explain the computational mechanisms that undergird attention based on neuroscience findings. We focused on reclaiming saliency as an active process that relies on neural gain control, uncertainty minimization and structured scheduling. Here, we describe how we can mathematically realize some of these mechanisms in the context of well-known challenges in robotics. Enabling robots with this type of attention may be crucial to filter the sensory signals and internal variables that are relevant to estimate the robot/world state and complete any task. More importantly, the active component of salience (i.e., behavior) is essential to interact with the world—as argued in active perception approaches (Bajcsy et al., 2018).

We revisit the standard view of attention in robotics by introducing sensory precision (inverse variance) as the driving mechanism for modulating both perception and action (Friston et al., 2011; Clark, 2013). Although saliency was originally described to underwrite behavior, most models used in robotics, strongly biased by computer vision approaches, focus on computing the most relevant region of an image (Borji and Itti, 2012)—mainly computing human fixation maps—relegating action to a secondary process. Illustratively, state-of-the-art deep learning saliency models—as shown in the MIT saliency benchmark (Bylinskii et al., 2019)—do not have the action as an output. Conversely, the active perception approach properly defines the action as an essential process of active sensing to gather the relevant information. Our proposed model, based on precision modulated action and perception coupling (*i*) place attention as essential for state-estimation and system identification and (*ii*) and reclaims saliency as a driver for information-seeking behavior, as proposed in early works (Tsotsos et al., 1995), but goes beyond human fixation maps for both improving the model of the environment (exploration) and solving the task (exploitation).

In what follows, we highlight the key role of precision by reviewing relevant brain-inspired attention models deployed in robotics (Section 4.1). We propose precision-modulated attentional mechanisms for robots in three contexts—perception (Section 4.2), action (Section 4.3) and active perception (Section 4.3.3). The precision-modulated perception is formalized for a robotics setting; *via* (*i*) state estimation (i.e., estimating the hidden states of a dynamic system from sensory signals—Section 4.2.2), and (*ii*) system identification (i.e., estimating the parameters of the dynamic system from sensory signals—Section 4.2.3). Next, we show

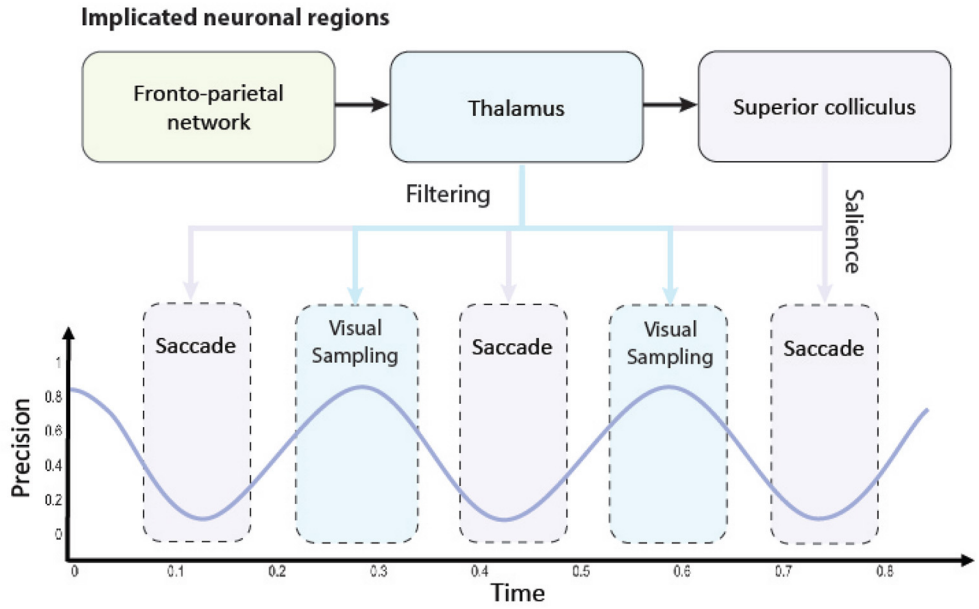



FIGURE 1
A graphical illustration of the precision-modulated account of perception and action. Salience and attention are computed based upon beliefs (assumed to be) encoded in parts of the fronto-parietal network and realized in distinct brain regions: superior colliculus (SC) for perception as inference and thalamus for planning as inference, respectively. To deploy attentional processes efficiently, these two mechanisms have to be aligned, which is done rhythmically, hypothetically in theta frequency. This coupling enables the saccadic suppression phenomenon through fluctuations in precision (on an arbitrary scale). When precision is low (i.e., the trough of the theta rhythm), the saccade emerges. Note that there might be distinct processes inhibiting the action (e.g., covert attention), and (despite a decline in precision) saccades might not emerge in every theta cycle. On the other hand, high precision facilitates confident inferences about the causes of visual data. Under this account, thalamus is used for initiating gain control (or visual sampling in general) by providing stronger sensory input, while superior colliculus dictates next saccades, that lead to most informative fixation positions.

TABLE 1 Robotics applications and their precision realizations.

| Task | Application | Precision manipulation | Sections |
|---|---|---|---|
| Perception | State and input estimation | Noise precision modeling $\tilde{\Pi}$ | 4.2.2 |
| | System Identification | Posterior parameter precision learning $\Pi^{\theta}$ | 4.2.3 |
| | Exploration-exploitation in learning | Prior parameter precision modeling $P^{\theta}$ | 4.2.4 |
| | Noise estimation | Noise precision learning $\tilde{\Pi}$ | 4.2.5 |
| Action | Informative Path Planning (IPP) | Precision optimization (of map) | 4.3.2 |
| Active perception | IPP with action-perception cycle | Precision modulation | 4.3.3 |

TABLE 2 Precision parameters that are manipulated in Section 4.2.

| Term | Symbol | Definition |
|---|---|---|
| Sensory precision | $\Pi^{z}$ | Inverse covariance of sensory noise $\mathbf{z}$ (Equation 1). |
| Prior parameter precision | $P^{\theta}$ | The robot's confidence on its prior parameters $\eta^{\theta}$. |
| Noise precision | $\tilde{\Pi}$ | The inverse covariance of all noises (Equation 5). |
| Posterior parameter precision | $\Pi^{\theta}$ | The robot's confidence on its parameter estimates. |

that precision-modulated action can be realized through precision optimization (planning future actions—Section 4.3.2) and discuss practical considerations for coupling with precision-modulated perception (precision based active perception—Section 4.3.3). Table 1 summarizes our proposed precision manipulations to solve relevant problems in robot perception and action. Table 2 provides the definitions of precision within our mechanism.

## 4.1. Previous brain-inspired attention models in robotics

Brain-inspired attention has been mainly addressed in robotics from a "passive" visual saliency perspective, e.g., which pixels of the image are the most relevant. This saliency map is then generally used to foveate the most salient region.

This approach was strongly influenced by early computational models of visual attention (Tsotsos et al., 1995; Itti and Koch, 2001). The first models deployed in robots were bottom-up, where the sensory input was transformed into an array of values that represents the importance (or salience) of each cue. Thus, the robot was able to identify which region of the scene has to look at, independently of the task performed—see Borji and Itti (2012) for a review on visual saliency. These models have also been useful for acquiring meaningful visual features in applications, such as object recognition (Orabona et al., 2005; Frintrop, 2006), localization, mapping and navigation (Frintrop and Jensfelt, 2008; Roberts et al., 2012; Kim and Eustice, 2013). Saliency computation was usually employed as a helper for the selection of the relevant characteristics of the environment to be encoded. Thus, reducing the information needed to process.

More refined methods of visual attention employed top-down modulation, where the context, task or goal bias the relevance of the visual input. These methods were used, for instance, to identify humans using motion patterns (Butko et al., 2008; Morén et al., 2008). A few works also focused on object/target search applications, where top-down and bottom-up saliency attention were used to find objects or people in a search and rescue scenario (Rasouli et al., 2020).

Attention has also been considered in human-robot interaction and social robotics applications (Ferreira and Dias, 2014), mainly for scene or task understanding (Kragic et al., 2005; Ude et al., 2005; Lanillos et al., 2016), and gaze estimation (Shon et al., 2005) and generation (Lanillos et al., 2015a). For instance, computing where the human is looking at and where the robot should look at or which object should be grasped. Furthermore, multi-sensory and 3D saliency computation has also been investigated (Lanillos et al., 2015b). Finally, more complex attention behaviors, particularly designed for social robotics and based on human non-verbal communication, such as joint attention, have also been addressed. Here the robot and the human share the attention of one object through meaningful saccades, i.e., head/eye movements (Nagai et al., 2003; Kaplan and Hafner, 2006; Lanillos et al., 2015a).

Although attention mechanisms have been widely investigated in robotics, specially to model visual cognition (Kragic et al., 2005; Begum and Karray, 2010), the majority of the works have treated attention as an extra feature that can help the visual processing, instead of a crucial component needed for the proper functioning of the cognitive abilities of the robot (Lanillos and Cheng, 2018a). Furthermore, these methods had the tendency to leave the action generation out of the attention process. One of the reasons for not including saliency computation, in robotic systems, is that the majority of the models only output "human-fixation map" predictions, given a static image. Saliency computation introduces extra computational complexity, which can be finessed by visual segmentation algorithms (e.g., line detectors in autonomous navigation). However, it does not resolve uncertainty nor select actions that maximize information gain in the future. In essence, the incomplete view of attention models that output human-fixation maps has arguably obscured the huge potential of neuroscience-inspired attentional mechanisms for robotics.

Our proposed model of attention, based on precision modulation, abandons the current robotics narrow view of attention and saliency by explicitly modeling attention within state estimation, learning and control. Thus, placing attentional processes at the core of the robot computation and not as an extra add-on. In the following sections, we describe the realization of our precision-based attention formulation in robotics using common practical applications as the backbone motif.

## 4.2. Precision-modulated perception

We formalize precision-modulated perception from a first principles Bayesian perspective—explicitly the free energy principle approach proposed by Friston et al. (2011). Practically, this entails optimizing precision parameters over (particular) model parameters.

Through numerical examples show how our model is able to perform accurate state estimation (Bos et al., 2021) and stable parameter learning (Meera and Wisse, 2021a,b). To illustrate the approach, we first introduce a dynamic system modeled as a linear state space system in robotics (Section 4.2.1)—we used this formulation in all our numerical experiments. We briefly review the formal terminologies for a robotics context to appropriately situate our precision-based mechanism for perception. Explicitly, we introduce: precision modeling (by adapting a known form of the precision matrix), precision learning (by learning the full precision matrix), and precision optimization (use precision as an objective function during learning). As a reminder, precision modeling is associated with (instantaneous) gain control and precision learning (at slower time scales) is associated with optimizing that control.

### 4.2.1. Precision for state space models

A linear dynamic system can be modeled using the following state space equations (boldface notation denotes components of the real system and non-boldface notation its estimates):

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{w}, \qquad \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{z}. \tag{1}$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are constant matrices defining the system parameters, $\mathbf{x} \in \mathbb{R}^n$ is the system state (usually an unobserved variable), $\mathbf{u} \in \mathbb{R}^r$ is the input or control actions, $\mathbf{y} \in \mathbb{R}^m$ is the output or the sensory measurements, $\mathbf{w} \in \mathbb{R}^n$ is the process noise with precision $\Pi^{\mathbf{w}}$ (or inverse variance $\Sigma^{\mathbf{w}-1}$), and $\mathbf{z} \in \mathbb{R}^m$ is the measurement noise with precision $\Pi^{\mathbf{z}}$.

For instance, we can describe a mass-spring damper system (depicted in Figure 2B) using state space equations. A mass ($m = 1.4kg$) is attached to a spring with elasticity constant ($k = 0.8N/m$), and a damper with a damping coefficient ($b = 0.4Ns/m$). When a force ($u(t) = e^{-0.25(t-12)^2}$) is applied on the mass, it displaces $x$ from its equilibrium point. The linear dynamics of this system is given by:

$$\begin{bmatrix} \dot{x} \\ \ddot{x} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u, \qquad y = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \end{bmatrix}. \quad (2)$$

Note that Equation (2) is equivalent to Equation (1) with parameters $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0, \frac{1}{m} \end{bmatrix}^T$ and $\mathbf{C} = \begin{bmatrix} 1 & 0 \end{bmatrix}$, and state $\mathrm{x} = \begin{bmatrix} x, \dot{x} \end{bmatrix}^T$.

Now we introduce attention as precision modulation assuming that the robotic goal is to minimize the prediction error (Friston et al., 2011; Lanillos and Cheng, 2018b; Meera and Wisse, 2020), i.e., to refine its model of the environment and perform accurate state estimation, given the information available. In other words, the robot has to estimate $\mathbf{x}$ and $\mathbf{u}$ from input prior $\eta^u$ with a prior precision of $P^u$, given the measurements $\mathbf{y}$, parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and noise precision $\Pi^{\mathbf{w}}$ and $\Pi^{\mathbf{z}}$. Formally, the prediction error $\tilde{\epsilon}$ of the sensory measurements $\tilde{\epsilon}^y$, control input reference $\tilde{\epsilon}^u$ and state $\tilde{\epsilon}^x$ are:

$$\tilde{\epsilon} = \begin{bmatrix} \tilde{\epsilon}^y \\ \tilde{\epsilon}^u \\ \tilde{\epsilon}^x \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{y}} - \tilde{\mathbf{C}}\tilde{x} \\ \tilde{u} - \tilde{\eta}^u \\ D^x\tilde{x} - \tilde{\mathbf{A}}\tilde{x} - \tilde{\mathbf{B}}\tilde{u} \end{bmatrix} \begin{cases} \text{sensory prediction error} \\ \text{control input prediction} \\ \text{error} \\ \text{state prediction error} \end{cases} \quad (3)$$

equivalent to shifting up all the components in generalized coordinates by one block.

We can estimate the state and input using the Dynamic Expectation Maximization (DEM) algorithm (Friston et al., 2008; Meera and Wisse, 2020) that optimizes a free energy variational bound $\mathcal{F}$ to be tractable[4]. This is:

$$X = \begin{bmatrix} \tilde{x} \\ \tilde{u} \end{bmatrix} = \arg\max_X \mathcal{F} = \arg\max_X -\frac{1}{2}\tilde{\epsilon}^T\tilde{\Pi}\tilde{\epsilon} \quad (4)$$

Crucially, $\tilde{\Pi}$ is the generalized noise precision that modulates the contribution of each prediction error to the estimation of the state and the computation of the action. Thus, $\tilde{\Pi}$ is equivalent to attentional gain. For instance, we can model the precision matrix to attend to the most informative signal derivatives in $\tilde{\mathbf{y}}$. Concisely, the precision $\tilde{\Pi}$ has the following form:

$$\tilde{\Pi} = \begin{bmatrix} S \otimes \Pi^z & 0 & 0 \\ 0 & S \otimes P^u & 0 \\ 0 & 0 & S \otimes \Pi^w \end{bmatrix}, \quad (5)$$

where $S$ is the smoothness matrix. In Section 4.2.2, we show that modeling the precision matrix $\tilde{\Pi}$ using the $S$ matrix improves the estimation quality.

The full free energy functional (time integral of free energy $\bar{\mathcal{F}} = \int \mathcal{F} dt$ at optimal precision) that the robot optimizes to perform state-estimation and system identification is described in Equation (6)—for readability we omitted the details of the derivation of this cost function, and we refer to Anil Meera and Wisse (2021) for further details.

$$\bar{\mathcal{F}} = -\frac{1}{2}\sum_t \Big[ \underbrace{\tilde{\epsilon}^{yT}\tilde{\Pi}^z\tilde{\epsilon}^y + \tilde{\epsilon}^{uT}P^{\tilde{u}}\tilde{\epsilon}^u + \tilde{\epsilon}^{xT}\tilde{\Pi}^w\tilde{\epsilon}^x}_{\text{precision weighed prediction error}} \Big] - \frac{1}{2}\Big[ \underbrace{\epsilon^{\theta T}P^\theta\epsilon^\theta + \epsilon^{\lambda T}P^\lambda\epsilon^\lambda}_{\text{prior precision weighed prediction error of } \theta \text{ and } \lambda} \Big]$$
$$+ \underbrace{\frac{1}{2}n_t \ln|\Sigma^X|}_{\text{state and input entropy}} + \frac{1}{2}n^t \underbrace{\Big[\ln|\tilde{\Pi}^z| + \ln|P^{\tilde{\nu}}| + \ln|\tilde{\Pi}^w|\Big]}_{\text{noise entropy}} + \underbrace{\frac{1}{2}\ln|\Sigma^\theta P^\theta|}_{\text{parameter entropy}} + \underbrace{\frac{1}{2}\ln|\Sigma^\lambda P^\lambda|}_{\text{hyperparameter entropy}} \quad (6)$$

Note that $\tilde{\epsilon}^y = \tilde{\mathbf{y}} - \tilde{\mathbf{C}}\tilde{x}$ is the difference between the observed measurement and the predicted sensory input given the state[3]. Here $D^x$ performs the (block) derivative operation, which is

Here $\epsilon^\theta = \theta - \eta^\theta$, $\epsilon^\lambda = \lambda - \eta^\lambda$ are the prediction errors of parameters and hyper-parameters[5]. $\bar{\mathcal{F}}$ consist of two main components: i) precision weighed prediction errors and ii)

3 The tilde over the variable refers to the generalized coordinates, i.e., the variable includes all temporal derivatives. Thus, $\tilde{\epsilon}$ is the combined prediction error of outputs, inputs and states. For example, the generalized output $\tilde{y}$ is given by $\tilde{y} = [y, y', y''...]^T$, where the prime operator denotes the derivatives. We use generalized coordinates (Friston et al., 2010) for achieving accurate state and input estimation during the presence of (colored) noise by modeling the time dependent quantities $(x, v, y, w, z)$ in generalized coordinates. This involves keeping track of the evolution of the trajectory of the probability distributions of states, instead of just their point estimates. Here the colored noise $w$ and $z$ are modeled as a white noise convoluted with a Gaussian kernel. The use of generalized coordinates has recently shown to outperform classical approaches under colored noise on real quadrotor flight (Bos et al., 2021).

4 Note that this expression of the variational free energy is using the Laplace and mean-field approximations commonly used in the FEP literature.

5 System identification involves the estimation of system parameters (denoted by $\theta$, e.g., vectorised A), given y, u, by starting from a parameter prior of $\eta^\theta$ with prior precision $P^\theta$, and a prior on noise hyper-parameter $\eta^\lambda$ with a prior precision of $P^\lambda$. Note that we parametrise noise precision

precision-based entropy. The dominant role of precision—in the free energy objective—is reflected in how modulating these precision parameters can have a profound influence perception and behavior. The theoretical guarantees for stable estimation (Meera and Wisse, 2021b), and its application on real robots (Lanillos et al., 2021) make this formulation very appealing to robotic systems.

Note that we can manipulate three kinds of precision within the state space formulation: (i) prior precision ($P^{\tilde{u}}, P^{\theta}, P^{\lambda}$), (ii) conditional precision on estimates ($\Pi^{X}, \Pi^{\theta}, \Pi^{\lambda}$) and (iii) noise precision ($\Pi^{z}, \Pi^{w}$). Therefore, to learn the correct parameter values $\theta$, we (i) learn the parameter precision $\Pi^{\theta}$, (ii) model the prior parameter precision $P^{\theta}$, and (iii) learn the noise precision $\Pi^{w}$ and $\Pi^{z}$ (parameterised using $\lambda$).

### 4.2.2. State and input estimation

State estimation is the process of estimating the unobserved states of a real system from (noisy) measurements. Here, we show how we can achieve accurate estimation through precision modulation in a linear time invariant system under the influence of colored noise (Meera and Wisse, 2020). State estimation in the presence of colored noise is inherently challenging, owing to the non-white nature of the noise, which is often ignored in conventional approaches, such as the Kalman Filter (Welch and Bishop, 2002).

Figure 2 summarizes a numerical example that shows how one can use precision modulation to focus on the less noisy derivatives (lower derivatives) of measurements, relative to imprecise higher derivatives. Thus, enabling the robot to use the most informative data for state and input estimation, while discarding imprecise input. Figure 2B depicts the mass-spring damper system used. The numerical results show that the quality of the estimation increases as the embedding ordering increases but the lack of information in the higher order derivatives of the sensory input do not affect the final performance due to the precision modulation. The higher order derivatives (Figure 2A) are less precise than the lower derivatives, thereby reflecting the loss of information in higher derivatives. The state and input estimation was performed using the optimization framework described in the previous section. The quality of estimation is shown in Figure 2C, where the input estimation using six derivatives (blue curve) is closer to the real input (yellow curve) than when compared to the estimation using only one derivative (red curve). The quality of the estimation reports the sum of squared error (SSE) in the estimation of states and inputs with respect to the embedding order (number of signal derivatives considered).

To obtain accurate state estimation by optimizing the precision parameters, we recall that the precision weights the prediction errors. From Equation (3), the structural form of $\tilde{\Pi}$ is mainly dictated by the smoothness matrix $S$, which establishes the interdependence between the components of the variable expressed in generalized coordinates (e.g., the dependence between $\mathbf{y}$, $\mathbf{y}'$ and $\mathbf{y}''$ in $\tilde{\mathbf{y}}$). For instance, the $S$ matrix for a Gaussian kernel is as follows Meera and Wisse (2022):

$$S = \begin{bmatrix} \frac{35}{16} & 0 & \frac{35}{8}s^2 & 0 & \frac{7}{4}s^4 & 0 & \frac{1}{6}s^6 \\ 0 & \frac{35}{4}s^2 & 0 & 7s^4 & 0 & s^6 & 0 \\ \frac{35}{8}s^2 & 0 & \frac{77}{4}s^4 & 0 & \frac{19}{2}s^6 & 0 & s^8 \\ 0 & 7s^4 & 0 & 8s^6 & 0 & \frac{4}{3}s^8 & 0 \\ \frac{7}{4}s^4 & 0 & \frac{19}{2}s^6 & 0 & \frac{17}{3}s^8 & 0 & \frac{2}{3}s^{10} \\ 0 & s^6 & 0 & \frac{4}{3}s^8 & 0 & \frac{4}{15}s^{10} & 0 \\ \frac{1}{6}s^6 & 0 & s^8 & 0 & \frac{2}{3}s^{10} & 0 & \frac{4}{45}s^{12} \end{bmatrix}, \quad (7)$$

where $s$ is the kernel width of the Gaussian filter that is assumed to be responsible for serial correlations in measurement or state noise. Here, the order of generalized coordinates (number of derivatives under consideration) is taken as six ($S \in \mathbb{R}^{7\times7}$). For practical robotics applications, the measurement frequency is high, resulting in $0 < s < 1$. It can be observed that the diagonal elements of $S$ decreases because $s < 1$, resulting in a higher attention (or weighting) on the prediction errors from the lower derivatives when compared to the higher derivatives. The higher the noise color (i.e., $s$ increases), the higher the weight given to the higher state derivatives (last diagonal elements of $S$ increases). This reflects the fact that smooth fluctuations have more information content in their higher derivatives. Having established the potential importance of precision weighting in state estimation, we now turn to the estimation (i.e., learning) of precision in any given context.

### 4.2.3. System identification

This section shows how to optimize system identification by means of precision learning (Anil Meera and Wisse, 2021; Meera and Wisse, 2021b). Specifically, we show how to fuse prior knowledge about the dynamic model with the data to recover unknown parameters of the system through an attention mechanism. This involves the learning of the (1) parameters and (2) noise precisions. Our model "turns" the attention to the least precise parameters and uses the data to update those parameters to increase their precision. Hence, allowing faster parameter learning.

For the sake of clarity, we use again the mass-spring-damper system as the driving example (Section 4.2.1). We formalize system identification as evaluating the unknown parameters $k$, $m$ and $b$, given the input $\mathbf{u}$, the output $\mathbf{y}$, and the general form of the linear system in Equation (2).

Figure 3 depicts the process of learning unknown parameters (dotted boxes denote the processes inside the robot brain). The robot measures its position $x(t)$ using its sensors (e.g., vision or range sensor). We assume that the robot has observed the behavior of a mass-spring-damper system

($\Pi^{w}$ and $\Pi^{z}$) using $\lambda \in \mathbb{R}^{2\times1} = \begin{bmatrix} \lambda^{z} \\ \lambda^{w} \end{bmatrix}$ as an exponential relation (e.g., $\Pi^{w}(\lambda^{w}) = \exp(\lambda^{w})I^{n\times n}$).

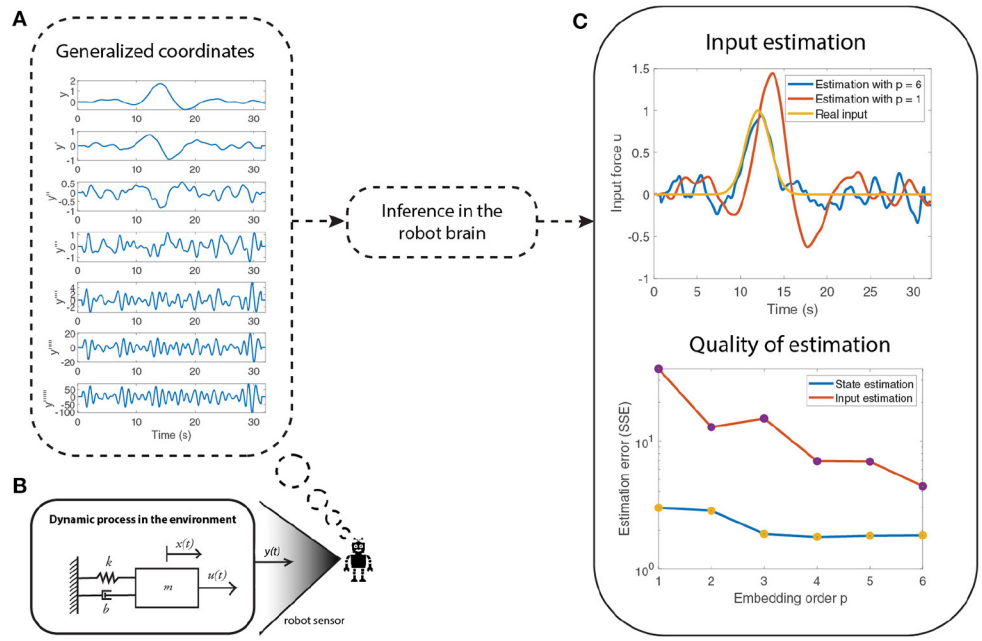



FIGURE 2
An illustration of an attention mechanism for state and input estimation of a system (shown in **B**). The quality of the estimation improves **(C)** as the embedding order (number of derivatives) of generalized coordinates are increased **(A)**. However, the imprecise information in the higher order derivatives of the sensory input **y** does not affect the final performance of the observer because of attentional selection, which selectively weighs the importance afforded to each derivative, in the free energy optimization scheme.

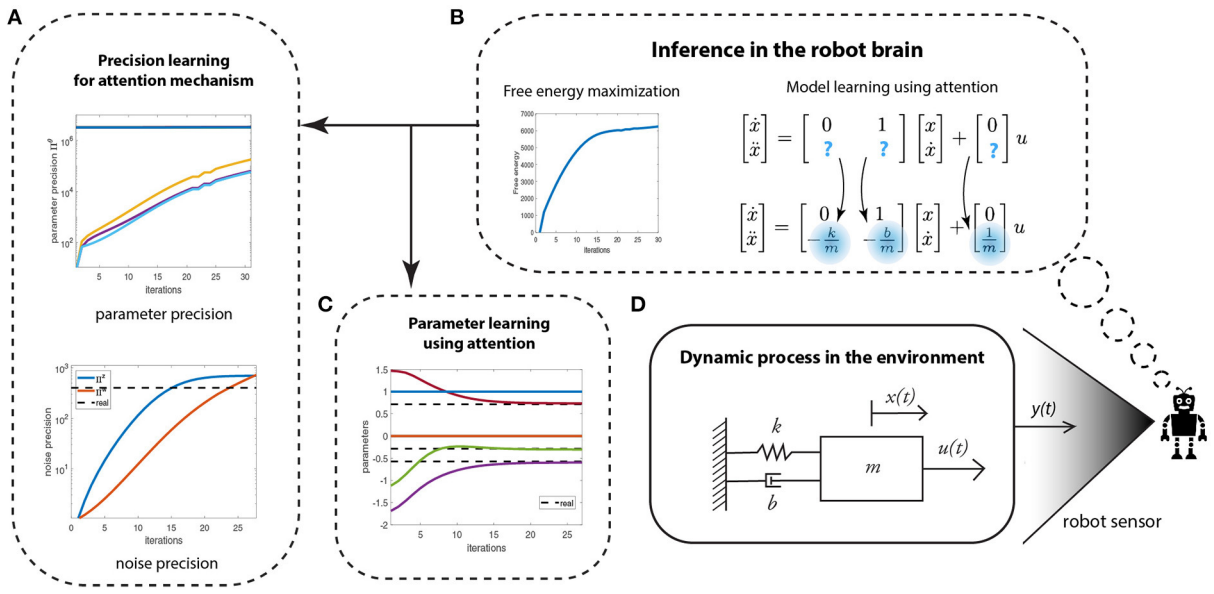



FIGURE 3
The schematic of the robot's attention mechanism for learning the least precise parameters of a given generative model of a mass-spring-damper system (shown in **D**). **(A)** Learning the conditional precision on parameters and the noise precision. **(B)** The free energy optimization helping to identify the unknown system parameters. **(C)** The parameter learning.

before or a model is provided by the expert designer. However, some of the parameters are unknown. The robot can reuse the prior learned model of the system to relearn the new system. This can be realized by setting a high prior precision on the known parameters and a low prior precision on the unknown parameters. By means of precision learning, the robot uses the

sensory signals to learn the parameter precision $\Pi^{\theta}$, thereby improving the confidence in the parameter estimates $\theta$. This directs the robot's attention toward the refinement of the parameters with least precision as they are the most uncertain. The requisite parameter learning proceeds by the gradient ascent of the free energy functional given in Equation (6). The parameter precision learning proceeds by tracking the negative curvature of $\bar{\mathcal{F}}$ as $\Pi^{\theta} = -\frac{\partial^2 \bar{\mathcal{F}}}{\partial \theta^2}$ (Anil Meera and Wisse, 2021).

The learning process—by means of variational free energy optimization (maximization)—is shown in Figure 3B. The learning involves two parallel processes: precision learning (Figure 3A), and parameter learning (Figure 3C). Precision learning comprises of parameter precision learning (top graph)—i.e., identifying the precision of an approximate posterior density for the parameters being estimated—and noise precision learning (bottom graph). The high prior precision on the known system parameters (0 and 1), and low prior precision on the unknown system parameters ($-\frac{k}{m}, -\frac{b}{m}$ and $\frac{1}{m}$, highlighted in blue) directs attention toward learning the unknown parameters and their precision. Note that in Figure 3A, the precision on the three unknown parameters start from a low prior precision of $P^{\theta} = 1$ and increase with each iteration, whereas the precision of known parameters (0 and 1) remains a constant ($3.3 \times 10^6$). The noise precisions are learned simultaneously, which starts from a low prior precision of $P^{\lambda^w} = P^{\lambda^z} = 1$ and finally converges to the true noise precision (dotted black line). Both precisions are used to learn the three parameters of the system (Figure 3B), which starts from randomly selected values within the range $[-2,2]$ and finally converges to the true parameter values of the system ($\theta_3 = -\frac{k}{m} = -0.5714$, $\theta_4 = -\frac{b}{m} = -0.2857$ and $\theta_6 = \frac{1}{m} = 0.7143$), denoted by black dotted lines. From an attentional perspective, the lower plot in Figure 3A is particularly significant here. This is because the robot discovers the data are more informative than initially assumed, thereby leading to an increase in its estimate of the precision of the data-generating process. This means that the robot is not only using the data to optimize its beliefs about states and parameters (system identification), it is also using these data to optimize the way in which it assimilates these data.

In summary, precision-based attention, in the form of precision learning, helps the robot to accurately learn unknown parameters by fusing prior knowledge with new incoming data (sensory measurements), and attending to the least precise parameters.

### 4.2.4. Precision-modulated exploration and exploitation in system identification

Exploration and exploitation in the parameter space can be advantageous to robots during system identification. Precision-based attention—here the prior precision—allows a graceful balance between the two, mediated by the prior precision[6]. A very high prior precision encourages exploitation and biases the robot toward believing its priors, while a low prior precision encourages exploration and makes the robot sensitive to new information.

We use again the mass-spring-damper system example but with a different prior parameter precision $P^{\theta}$. The prior parameters are initialized at random and learned using optimization. Figure 4B shows the increase in parameter estimation error (SSE) as the prior parameter precision $P^{\theta}$ increases until it finally saturates. The bottom left region (circled in red) indicates the region where the prior precision is low, encouraging exploration with high attention on the sensory signals for learning the model. This region over-exposes the robot to its sensory signals by neglecting the prior parameters. The top right region (circled in red) indicates the biased robot where the prior precision is high, encouraging the robot to exploit its prior beliefs by retaining high attention on prior parameters. This regime biases the robot into being confident about its priors and disregarding new information from the sensory signals. Between those extreme regimes (blue curve) the prior precision balances the exploration-exploitation trade-off. Figure 4A describes how increased attention to sensory signals helped the robot to recover from poor initial estimates of parameter values and converge toward the correct values (dotted black line). Conversely, in Figure 4C, high attention on prior parameters did not help the robot to learn the correct parameter values.

These results establish that prior precision modeling allows balanced exploration and exploitation of parameter space during system identification. Although the results show that an over-exposed robot provides better parameter learning, we show—in the next section—that this is not always be the case.

### 4.2.5. Noise estimation

In real-world applications, sensory measurements are often highly noisy and unpredictable. Furthermore, the robot does not have access to the noise levels. Thus, it needs to learn the noise precision ($\Pi^{\mathbf{z}}$) for accurate estimation and robust control. Precision-based attention enables this learning. In what follows, we show how one can estimate $\Pi^{\mathbf{z}}$ using noise precision learning and that biasing the robot to prior beliefs can be advantageous in highly noisy environments.

Consider again the mass-spring-damper system in Figure 5B, where heavy rainfall/snow corrupts visual sensory signals. We evaluate the parameter estimation error under different noise conditions, using different levels of noise

6 Note that here we are using exploration and exploration not in terms of behavior but for parameter learning. Exploration means adapting the parameter to a different (unexplored) value and exploitation means keeping that value.

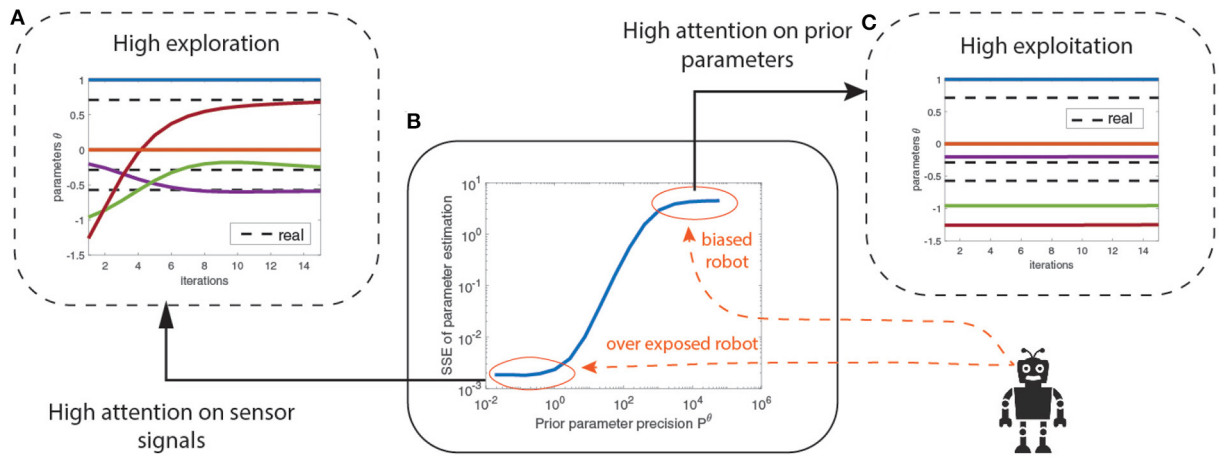



FIGURE 4
**(A)** Lower $P\hat{}\theta$ gives a high exploration strategy across the parameter space. **(B)** Precision-based attention allows exploration and exploitation balanced model learning mediated by the prior precisions on the parameters $P\hat{}\theta$. **(C)** The higher the $P\hat{}\theta$, the higher the attention on prior parameters $\eta\hat{}\theta$ and the lower the attention on the sensory signals while learning.

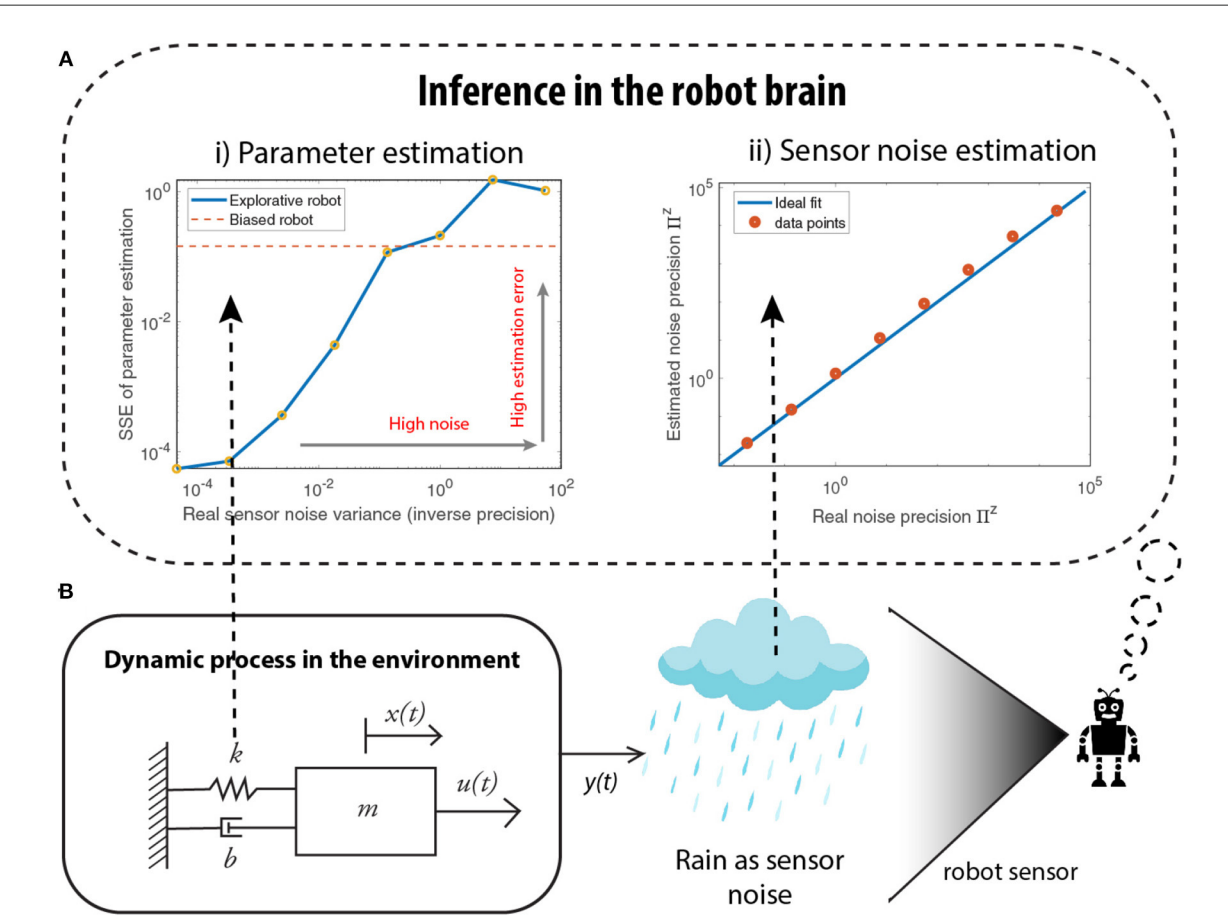



FIGURE 5
Simulations demonstrating how a biased robot could be advantageous, especially while learning in a highly noisy environment (shown in **B**). **(A i)** As the sensor noise increases, the quality of parameter estimation deteriorates to a point where an explorative robot generates higher parameter estimation errors than when compared to the biased robot that relies on its prior parameters. **(A ii)** However, the sensor noise estimation is accurate even for high noise environments, demonstrating the success of the attention mechanism using the noise precision.

variances (inverse precision). For an over-exposed robot (only attending to sensory measurements), left plot of Figure 5A, the estimation error increases as the noise strength increases, to a point where the error surpasses the error from a prior-biased robot. This shows that a robot, confident in its prior model, assigns low attention to sensory signals and outperforms an over-exposed robot that assigns high attention to sensory signals, in a highly noisy environment. The right plot of Figure 5A shows the quality of noise precision learning for an over-exposed robot. It can be seen that all the data points in red lie close to the blue line, indicating that the estimated noise precision is close to the real noise precision. Therefore, the robot is capable of recovering the correct sensory noise levels even when the environment is extremely noisy, where accurate parameter estimation is difficult.

These numerical results show that attention mechanism—by means of noise precision learning—allows the estimation of the noise levels in the environment and thereby protects against over-fitting or overconfident parameter estimation.

**Summary**. We have shown how precision-based attention–through precision modeling and learning– yields to accurate robot state estimation, parameter identification and sensory noise estimation. In the next section, we discuss how action is generated in this framework.

## 4.3. Precision-modulated action

Selecting the optimal sequence of actions to fulfill a task is essential for robotics (LaValle, 2006). One of the most prominent challenges is to ensure robust behavior given the uncertainty emerging from a highly complex and dynamic real world, where the robots have to operate on. A proper attention system should provide action plans that resolve uncertainty and maximize information gain. For instance, it may minimize the information entropy, thereby encouraging repeated sensory measurements (observations) on high uncertainty sensory information.

Salience, which in neuroscience is sometimes identified as Bayesian surprise (i.e., divergence between prior and posterior), describes which information is relevant to process. We go one step further by defining the saliency map as the epistemic value of a particular action (Friston et al., 2015). Thus, the (expected) divergence now becomes the mutual information under a particular action or plan. This makes the saliency map more sophisticated because it is an explicit measure of the reduction in uncertainty or mutual information associated with a particular action (i.e., active sampling), and more pragmatic because it tells you where to sample data next, given current Bayesian beliefs.

We first describe a precision representation usually used in information gathering problems and then how to directly generate action plans through precision optimization. Afterwards, we discuss the realization of the full-fledged model presented in the neuroscience section for active perception. We use the informative path planning (IPP) problem, described in Figure 6, as an illustrative example to drive intuitions.

### 4.3.1. Precision maps as saliency

One of the popular approaches in information gathering problems is to model the information map as a distribution [e.g., using Gaussian processes (Hitz et al., 2017)]. This is widely used in applications, such as a target search, coverage and navigation. The robot keeps track of an occupancy map and the associated uncertainty map (covariance matrix or inverse precision). While the occupancy map records the presence of the target on the map, the uncertainty map records the quality of those observations. The goal of the robot is to learn the distribution using some learning algorithm (Marchant and Ramos, 2014). A popular strategy is to plan the robot path such that it minimizes the uncertainty of the map in future (Popović et al., 2017). In Section 4.3.2, we will show how we can use the map precision to perform active perception, i.e., optimize the robot path for maximal information gain. Optimizing the map precision drives the robot toward an exploratory behavior.

### 4.3.2. Precision optimization for action planning

To introduce precision-based saliency we use an exemplary application of search and rescue. The goal is to find all humans using an unmanned air vehicle (UAV) (Lanillos, 2013; Lanillos et al., 2014; Meera et al., 2019; Rasouli et al., 2020). We use precision for two purposes: (i) precision optimization for action planning (plan flight path) and (ii) precision learning for map refinement. In contrast to previous models of action selection within active inference in robotics (Lanillos et al., 2021; Oliver et al., 2021) here precision explicitly drives the agent behavior. Figure 7 describes the scenario in simulation. The seven human targets on the ground are correctly identified by the UAV. We can formalize the solution as the UAV actions (next flight path) that minimize the future uncertainties of the human occupancy map. In our precision-based attention scheme, this objective is equivalent to maximizing the posterior precision of the map. Figure 8 shows the reduction in map uncertainty after subsequent assimilation of the measurements (camera images from the UAV, processed by a human detector). The map (and precision) is learned using a recursive Kalman Filter by fusing the human detector outcome onto the map (and precision). The algorithm drives the UAV toward the least explored regions in the environment, defined by the precision map.

Furthermore, Figure 9 shows an example of uncertainty resolution under false positives. In this case, human targets are moved to the bottom half of the map. The first measurement provides a wrong human detection with high uncertainty. However, after repeated measurements at the same location in

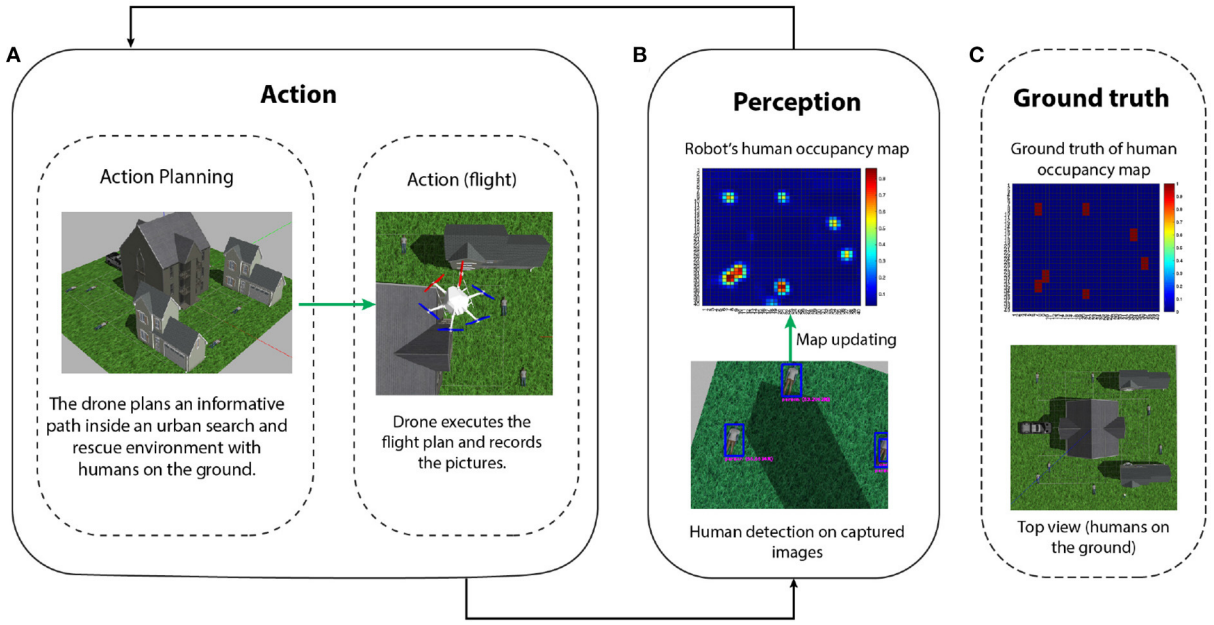



FIGURE 6
IPP problem for localizing human victims in an urban search and rescue scenario (Meera et al., 2019). **(A)** Action: a UAV, in a realistic simulation environment, plans a finite look-ahead path to minimize the uncertainty of its human occupancy map (e.g., modeled as a Gaussian process) of the world. The planned path is then executed, during which the UAV flies and captures images at a constant measurement frequency. **(B)** Perception: after the data acquisition is complete, a human detection algorithm is executed to detect all the humans on the images. These detections are then fused into the UAV's human location map. The cycle is repeated until the uncertainty of the map is completely resolved (this usually implies enough area coverage and repeated measurements on uncertain locations). The ground truth of the human occupancy map and the UAV belief is shown in **(B,C)** respectively. The final map approaches the ground truth and all the seven humans on the ground are correctly detected.

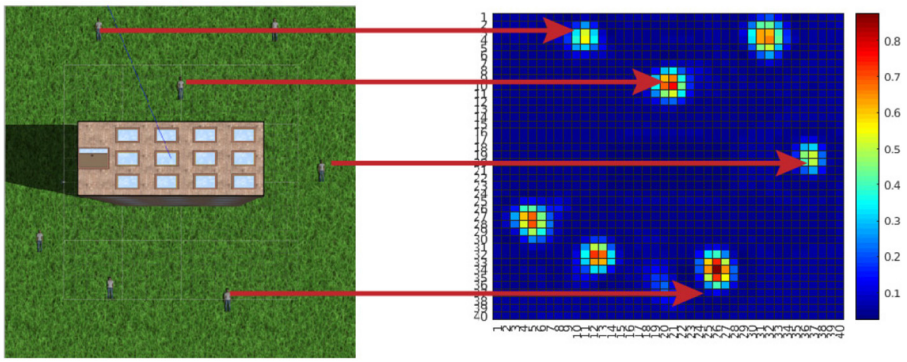



FIGURE 7
Finding humans with unmanned air vehicles (UAVs): an informative path planning (IPP) approach (Anil Meera, 2018). The simulation environment on the left consists of a tall building at the center, surrounded by seven humans lying on the floor. The goal of the UAV is to compute the action sequence that allows maximum information gathering, i.e., the humans location uncertainty is minimized. On the right is the final occupancy map colored with the probability of finding a human at that location. It can be observed that all humans on the simulation environment were correctly detected by the robot.

the map the algorithm was capable of resolving this ambiguity, to finally learn the correct ground truth map. Hence, the sought behavior is to take actions that encourage repeated measurements at uncertain locations for reducing uncertainty.

Although the IPP example illustrates how to generate control actions through precision optimization, the task, by construction, is constrained to explicitly reduce uncertainty. This is similar to the description of visual search described in Friston et al. (2012), where the location was chosen maximize information gain. Information gain (i.e., the Bayesian surprise expected following an action) is a key part of the expected free energy functional that underwrite action selection in active

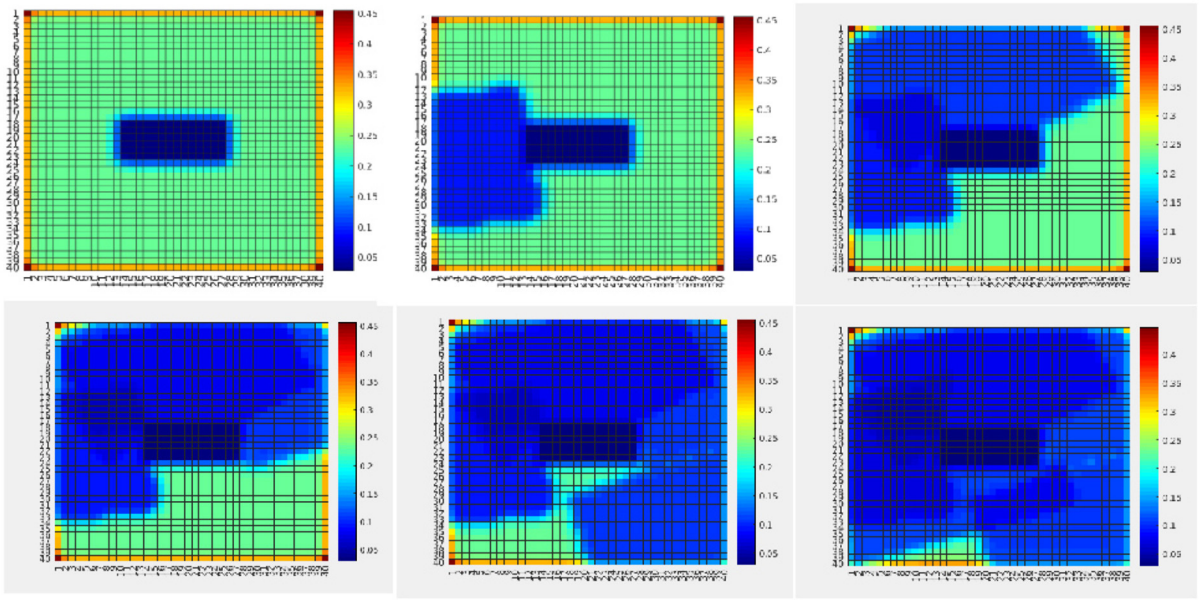


FIGURE 8
Variance map of the probability distribution of people location (Figure 7)—inverse precision of human occupancy map. The plot sequence shows the reduction of map uncertainty (inverse precision) after measurements (Anil Meera, 2018).

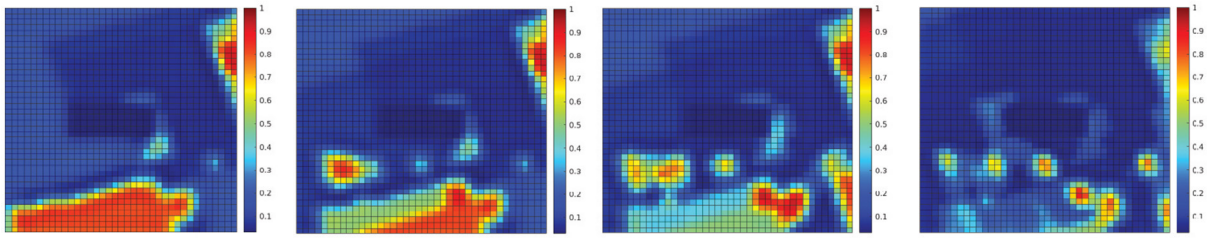


FIGURE 9
The human occupancy map (probability to find humans at every location of the environment) at four time instances during the UAV flight showing ambiguity resolution. The ambiguity arising from imprecise sensor measurements (false positive) is resolved through repeated measurements at the same location. The plot sequence shows how the assimilation of the measurements updates the probability of the people being in each location of the map (Meera et al., 2019).

inference. In brief, expected free energy can be decomposed into two parts the first corresponds to the information gain above (a.k.a., epistemic value or affordance). The second corresponds to the expected log evidence or marginal likelihood of sensory samples (a.k.a., pragmatic value). When this likelihood is read as a prior preference, it contextualizes the imperative to reduce uncertainty by including a goal-directed, imperative. For example, in the search paradigm above, we could have formulated the problem in terms of reducing uncertainty about whether each location was occupied by a human or not. We could have then equipped the agent with prior preferences for observing humans.

In principle, this would have produced searching behavior until uncertainty had been resolved about the scene; after which, the robot would seek out humans; simply because, these are its preferred outcomes. In thinking about how this kind of neuroscience inspired or biomimetic approach could be implemented in robotics, one has to consider carefully, the precision afforded sensory inputs (i.e., the likelihood of sensory data, given its latent causes)—and how this changes during robotic flight and periods of data gathering. This brings us back to the precision modulation and the temporal scheduling of searching and securing data. In the final section, we conclude with a brief discussion of how this might be implemented in future applications.

### 4.3.3. Precision-based active perception

In this section, we discuss the realization of a biomimetic brain-inspired model in relation to existing solutions in

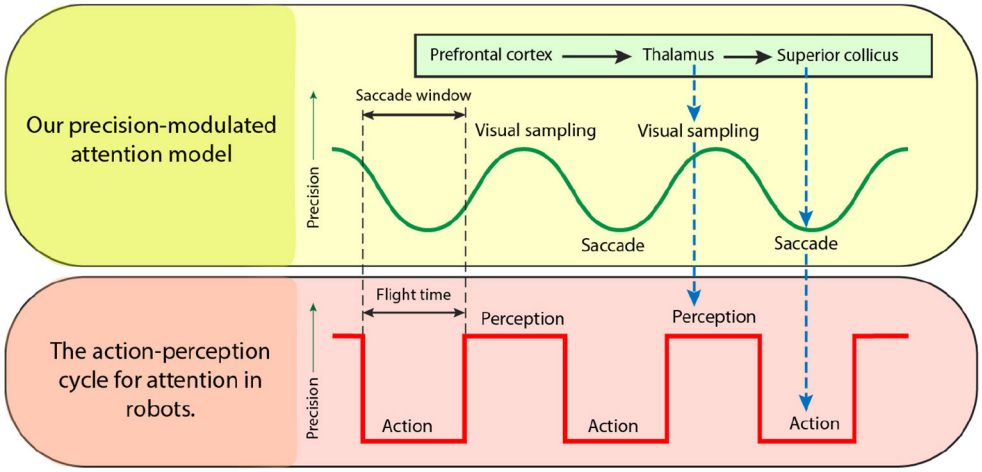



FIGURE 10
Precision-modulated attention model adapted to the action-perception loop in robotics. Each cycle consists of two steps: (1) action (planning and execution of a finite-time look ahead of the robot path for data collection) and (2) perception (learning using the collected data). This scheduling, using a finite time look-ahead plan, is quite common in real applications and of particular importance when processing is computationally expensive, e.g., slow rate of classification, non-scalable data fusion algorithms, Exponential planners, etc. However, the benefits of incorporating "optimal" scheduled loop driven by precision should be further studied.

robotics in the context of path-planning. Figure 10 compares our proposed precision-modulated attention model—from Figure 1—with the action-perception loop widely used in robotics. By analogy with eye saccades to the next visual sample, the UAV flies (action) over the environment to assimilate sensory data for an informed scene construction (perception). Once the flight time of the UAV is exhausted (similar to saccade window of the eye), the action is complete, after which the map is updated, and the next flight path is planned.

In standard applications of active inference, the information gain is supplemented with expected log preferences to provide a complete expected free energy functional (Sajid et al., 2021a). This accommodates the two kinds of uncertainty that actions and choices typically reduce. The first kind of uncertainty is inherent in unknowns in the environment. This is the information gain we have focused on above. The second kind of uncertainty corresponds to expected surprise, where surprise rests upon a priori expected or preferred outcomes. As noted above, equipping robots with both epistemic and pragmatic aspects to their action selection or planning could produce realistic and useful behavior that automatically resolves the exploration-exploitation dilemma. This follows because the expected free energy contains the optical mixture of epistemic (information-seeking) and pragmatic (i.e., preference seeking) components. Usually, after a period of exploration, the preference seeking components predominate because uncertainty has been resolved. Although expected free energy provides a fairly universal objective function for sentient behavior, it does not specify how to deploy behavior and sensory processing optimally. This brings us to the precision modulation model, inspired by neuroscientific considerations of attention and salience.

Hence, there are key differences between biological and robotic implementations of the search behavior. First, the use of oscillatory precision to modulate visual sampling and movement cycles, as opposed to arbitrary discrete action and perception steps currently used in robotics. Second, precision modulation influences both state estimation and action following the same uncertainty reduction principle. Importantly, our salience formulation speaks to selecting future data that reduces this uncertainty. For instance, we have shown—in the information gathering IPP example described in the previous subsection—that by optimizing precision we also optimize behavior.

Hence, there are key differences between biological and robotic implementations of the search behavior. First, the use of oscillatory precision to modulate visual sampling and movement cycles, as opposed to arbitrary discrete action and perception steps currently used in robotics. Second, precision modulation influences both state estimation and action following the same uncertainty reduction principle. Importantly, our salience formulation speaks to selecting future data that reduces this uncertainty. For instance, we have shown—in the information gathering IPP example described in the previous subsection—that by optimizing precision we also optimize behavior.

We argue the potential need and the advantages of realizing precision based temporal scheduling, as described the our brain-inspired model, for two practically relevant test cases: (*i*) learning dynamic models and (*ii*) information seeking applications.

In Section 4.2.4, we have shown how the exploration-exploitation trade-off can be mediated by the prior parameter

precision during learning. However, the accuracy-precision curve (Figure 4B) is often practically unavailable due to unknown true parameters values, challenging the modeling of prior precision. An alternative would be to use a precision based temporal scheduling mechanism to alternate between exploration and exploitation by means of a varying $P^{\theta}$ (similar to Figure 10) during learning, such that system identification is neither biased nor over exposed to sensory measurements. In Figure 5A, we showed how noise levels influence estimation accuracy, and how biasing the robot by modeling $P^{\theta}$ can be beneficial for highly noisy environments. A precision based temporal scheduling mechanism by means of a varying $P^{\theta}$ could provide a balanced solution between a biased robot (that exploits its model) and an exploratory one.

Furthermore, temporal scheduling, in the same way that eye saccades are generated, can be adapted for information gathering applications, such as target search, simultaneous localization and mapping, environment monitoring, etc. For instance, introducing precision-modulation scheduling for solving the IPP, and scheduling perception (map learning) and action (UAV flight). Precision modulation will switch between action and perception: when the precision is high, perception occurs (c.f., visual sampling), and when the precision is low, action occurs (c.f., eye movements). This switch, which is often implemented in the robotics literature using a budget for flight time, will be now dictated by precision dynamics.

In short, we have sketched the basis for a future realization of precision-based active perception, where the robot computes the actions to minimize the expected uncertainty. While most attentional mechanisms in robotics are limited to providing a "saliency" map highlighting the most relevant features, our attention mechanism proposes a general scheduling mechanism with action in the loop with perception, both driven by precision.

## 5. Concluding remarks

We have considered attention and salience as two distinct processes that rest upon oscillatory precision control processes. Accordingly, they require particular temporal considerations: attention to reliably estimate latent states from current sensory data and salience for uncertainty reduction regarding future data samples. This formulation addresses visual search from a first principles (Bayesian) account of how these mechanisms might manifest—and the circular causality that undergirds them *via* a rhythmic theta-coupling. Crucially, we have revisited the definition of salience from the visual neurosciences; where it is read as Bayesian surprise (i.e., the Kullback Leibler divergence between prior and posterior beliefs). We took this one step further and defined salience as the expected Bayesian surprise (i.e., epistemic value) of a particular action (e.g., sampling this set of data) (Friston et al., 2017b; Sajid et al., 2021a). Formulating salience as the expected divergence renders it the mutual information under a particular action (or action trajectory) (Friston et al., 2021),—and highlights its role in encoding working memory (Parr and Friston, 2017b). For brevity, our narrative was centered around visual attention and its realization *via* eye movements. However, this model does not strictly need to be limited to visual information processing, because it addresses sensorimotor and auditory processing in general. This means it explains how action and perception can be coupled in other sensory modalities. For instance, Tomassini et al. (2017) showed that visual information is coupled with finger movements at a theta rhythm.

The point of contact with the robotics use of salience emerges because the co-variation between a particular parameterisation and the inputs is a measure of the mutual information between the data and its estimated causes. In this sense, both definitions of salience reflect the mutual information—or information about a particular representation of a (latent) cause—afforded by an observation or consequence. However, our formulation is more sophisticated. Briefly, because it is an explicit measure of the reduction in uncertainty (i.e., mutual information) associated with a particular action (i.e., active sampling) and specifies where to sample data next, given current Bayesian beliefs. These processes (attention and salience) are a consequence of precision of beliefs over distinct model parameters. Explicitly, attention contends with precision over the causes of (current) outcomes and salience contends with beliefs about the data that has to be acquired and precision over beliefs about actions that dictate it. Since both processes can be linked *via* precision manipulation, the crucial thing is the precision that differentiates whether the agent acquires new information (under high precision) or resolves uncertainty by moving (low precision).

The focus of this work has been to illustrate the importance of optimizing precision at various places in generative models used for data assimilation, system identification and active sensing. A key point—implicit in these demonstrations - rests upon the mean field approximation used in all applications. Crucially, this means that getting the precision right matters, because updating posterior estimates of states, parameters and precisions all depend upon each other. This may be particularly prescient for making the most sense of samples that maximizes information gain. In other words, although attention and salience are separable optimization processes, they depend upon each other during active sensing. This was the focus of our final numerical studies of action planning.

To face-validate our formulation, we evaluated precision-modulated attentional processes in the robotic domain. We presented numerical examples to show how precision manipulation underwrites accurate state and noise estimation (e.g., selecting relevant information), as well as allowing system identification (e.g., learning unknown parameters of the dynamics). We also showed how one can use precision-based optimization to solve interesting problems; like the informative

path planning in search and rescue scenarios. Thus, in contrast to previous uses of attention in robotics, we placed attention and saliency as integral processes for efficient gathering and processing of sensory information. Accordingly, 'attention' is not only about filtering the current flow of information from the sensors but performing those actions that minimize expected uncertainty. Still, the full potential of our proposal has yet to be realized, as the precision-based attention should be able to account for prior preferences beyond the IPP problem (e.g., localizing people using UAVs). Finally, we briefly considered the realization of temporal scheduling for information gathering tasks, opening up interesting lines of research to provide robots with biologically plausible attention.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## Author contributions

AA and FN are responsible for the novel account and its translation to robotics. AA, FN, PL, and NS wrote the manuscript. All authors contributed to conception and design of the work, manuscript revision, read, and approved the submitted version.

## Funding

The open access publication of the manuscript was funded by the TU Delft Library. FN was funded by the Serotonin & Beyond project (953327). PL was partially supported by Spikeference project, Human Brain Project Specific Grant Agreement 3 (ID: 945539). NS was funded by the Medical Research Council (MR/S502522/1) and 2021-2022 Microsoft Ph.D. Fellowship. KF is supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z) and a Canada UK Artificial Intelligence Initiative (Ref: ES/T01279X/1).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Ahnelt, P. (1998). The photoreceptor mosaic. *Eye* 12, 531–540. doi: 10.1038/eye.1998.142

Anil Meera, A. (2018). *Informative path planning for search and rescue using a uav* (Maters thesis). TU Delft.

Anil Meera, A., and Wisse, M. (2021). Dynamic expectation maximization algorithm for estimation of linear systems with colored noise. *Entropy* 23, 1306. doi: 10.3390/e23101306

Atrey, A., Clary, K., and Jensen, D. (2019). Exploratory not explanatory: counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv[preprint].arXiv:1912.05743*. doi: 10.48550/arXiv.1912.05743

Axmacher, N., Henseler, M. M., Jensen, O., Weinreich, I., Elger, C. E., and Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3228–3233. doi: 10.1073/pnas.0911531107

Bajcsy, R., Aloimonos, Y., and Tsotsos, J. K. (2018). Revisiting active perception. *Auton. Robots* 42, 177–196. doi: 10.1007/s10514-017-9615-3

Balestrieri, E., Ronconi, L., and Melcher, D. (2021). Shared resources between visual attention and visual working memory are allocated through rhythmic sampling. *Eur. J. Neurosci.* 55, 3040–3053. doi: 10.1111/EJN.15264/v2/response1

Begum, M., and Karray, F. (2010). Visual attention for robotic cognition: a survey. *IEEE Trans. Auton. Ment. Dev.* 3, 92–105. doi: 10.1109/TAMD.2010.2096505

Benedetto, A., and Morrone, M. C. (2017). Saccadic suppression is embedded within extended oscillatory modulation of sensitivity. *J. Neurosci.* 37, 3661–3670. doi: 10.1523/JNEUROSCI.2390-16.2016

Benedetto, A., Morrone, M. C., and Tomassini, A. (2020). The common rhythm of action and perception. *J. Cogn. Neurosci.* 32, 187–200. doi: 10.1162/jocn_a_01436

Borji, A., and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Bos, F., Meera, A. A., Benders, D., and Wisse, M. (2021). Free energy principle for state and input estimation of a quadcopter flying in wind. *arXiv[preprint].arXiv:2109.12052*. doi: 10.48550/arXiv.2109.12052

Brown, H., Adams, R. A., Parees, I., Edwards, M., and Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cogn. Process.* 14, 411–427. doi: 10.1007/s10339-013-0571-3

Brzezicka, A., Kamiński, J., Reed, C. M., Chung, J. M., Mamelak, A. N., and Rutishauser, U. (2019). Working memory load-related theta power decreases in dorsolateral prefrontal cortex predict individual differences in performance. *J. Cogn. Neurosci.* 31, 1290–1307. doi: 10.1162/jocn_a_01417

Busch, N. A., and VanRullen, R. (2010). Spontaneous eeg oscillations reveal periodic sampling of visual attention. *Proc. Natl. Acad. Sci. U.S.A.* 107, 16048–16053. doi: 10.1073/pnas.1004801107

Butko, N. J., Zhang, L., Cottrell, G. W., and Movellan, J. R. (2008). "Visual saliency model for robot cameras," in *2008 IEEE International Conference on Robotics and Automation* (Pasadena, CA: IEEE), 2398–2403.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., et al. (2019). *Mit Saliency Benchmark*. Available online at: http://saliency.mit.edu/

Clark, A. (2013). The many faces of precision (replies to commentaries on "whatever next? neural prediction, situated agents, and the future of cognitive science"). *Front. Psychol.* 4, 270. doi: 10.3389/fpsyg.2013.00270

Crevecoeur, F., and Kording, K. P. (2017). Saccadic suppression as a perceptual consequence of efficient sensorimotor estimation. *Elife* 6, e25073. doi: 10.7554/eLife.25073

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *J. Math. Psychol.* 99, 102447. doi: 10.1016/j.jmp.2020.102447

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13494–13499. doi: 10.1073/pnas.93.24.13494

Dugué, L., Marque, P., and VanRullen, R. (2015). Theta oscillations modulate attentional search performance periodically. *J. Cogn. Neurosci.* 27, 945–958. doi: 10.1162/jocn_a_00755

Dugué, L., Roberts, M., and Carrasco, M. (2016). Attention reorients periodically. *Curr. Biol.* 26, 1595–1601. doi: 10.1016/j.cub.2016.04.046

Eldar, E., Cohen, J. D., and Niv, Y. (2013). The effects of neural gain on attention and learning. *Nat Neurosci.* 16, 1146–1153. doi: 10.1038/nn.3428

Feldman, H., and Friston, K. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4, 215. doi: 10.3389/fnhum.2010.00215

Ferreira, J. F., and Dias, J. (2014). Attentional mechanisms for socially interactive robots-a survey. *IEEE Trans. Auton. Ment. Dev.* 6, 110–125. doi: 10.1109/TAMD.2014.2303072

Fiebelkorn, I. C., and Kastner, S. (2019). A rhythmic theory of attention. *Trends Cogn. Sci.* 23, 87–101. doi: 10.1016/j.tics.2018.11.009

Fiebelkorn, I. C., and Kastner, S. (2020). Functional specialization in the attention network. *Annu. Rev. Psychol.* 71, 221–249. doi: 10.1146/annurev-psych-010418-103429

Fiebelkorn, I. C., and Kastner, S. (2021). Spike timing in the attention network predicts behavioral outcome prior to target selection. *Neuron* 109, 177–188. doi: 10.1016/j.neuron.2020.09.039

Fiebelkorn, I. C., Pinsk, M. A., and Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. *Neuron* 99, 842–853. doi: 10.1016/j.neuron.2018.07.038

Fiebelkorn, I. C., Pinsk, M. A., and Kastner, S. (2019). The mediodorsal pulvinar coordinates the macaque fronto-parietal network during rhythmic spatial attention. *Nat. Commun.* 10, 1–15. doi: 10.1038/s41467-018-08151-4

Fine, M. S., and Minnery, B. S. (2009). Visual salience affects performance in a working memory task. *J. Neurosci.* 29, 8016–8021. doi: 10.1523/JNEUROSCI.5503-08.2009

Frintrop, S., and Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual slam. *IEEE Trans. Rob.* 24, 1054–1065. doi: 10.1109/TRO.2008.2004977

Frintrop, S. (2006). *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, Vol. 3899*. Berlin: Springer.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Adams, R., Perrinet, L., and Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Front. Psychol.* 3, 151. doi: 10.3389/fpsyg.2012.00151

Friston, K., Da Costa, L., Hafner, D., Hesp, C., and Parr, T. (2021). Sophisticated inference. *Neural Comput.* 33, 713–763. doi: 10.1162/neco_a_01351

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Mattout, J., and Kilner, J. (2011). Action understanding and active inference. *Biol. Cybern.* 104, 137–160. doi: 10.1007/s00422-011-0424-z

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K., Stephan, K., Li, B., and Daunizeau, J. (2010). Generalised filtering. *Math. Problems Eng.* 2010, 621670. doi: 10.1155/2010/621670

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., and Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neurosci. Biobehav. Rev.* 118, 42–64. doi: 10.1016/j.neubiorev.2020.07.005

Friston, K. J., Trujillo-Barreto, N., and Daunizeau, J. (2008). Dem: a variational treatment of dynamic systems. *Neuroimage* 41, 849–885. doi: 10.1016/j.neuroimage.2008.02.054

Gazzaley, A., and Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16, 129–135. doi: 10.1016/j.tics.2011.11.014

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* 9, 181–197. doi: 10.1017/S0952523800009640

Helfrich, R. F., Fiebelkorn, I. C., Szczepanski, S. M., Lin, J. J., Parvizi, J., Knight, R. T., et al. (2018). Neural mechanisms of sustained attention are rhythmic. *Neuron* 99, 854–865. doi: 10.1016/j.neuron.2018.07.032

Helmholtz, H. V. (1925). *Treatise on Physiological Optics*. Rochester, NY: Optical Society of America.

Hitz, G., Galceran, E., Garneau, M.-È., Pomerleau, F., and Siegwart, R. (2017). Adaptive continuous-space informative path planning for online environmental monitoring. *J. Field Rob.* 34, 1427–1449. doi: 10.1002/rob.21722

Hogendoorn, H. (2016). Voluntary saccadic eye movements ride the attentional rhythm. *J. Cogn. Neurosci.* 28, 1625–1635. doi: 10.1162/jocn_a_00986

Hsieh, L.-T., and Ranganath, C. (2014). Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. *Neuroimage* 85, 721–729. doi: 10.1016/j.neuroimage.2013.08.003

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500

Kanai, R., Komura, Y., Shipp, S., and Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140169. doi: 10.1098/rstb.2014.0169

Kanwisher, N., and Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nat. Rev. Neurosci.* 1, 91–100. doi: 10.1038/35039043

Kaplan, F., and Hafner, V. V. (2006). The challenges of joint attention. *Interact. Stud.* 7, 135–169. doi: 10.1075/is.7.2.04kap

Kim, A., and Eustice, R. M. (2013). Real-time visual slam for autonomous underwater hull inspection using visual saliency. *IEEE Trans. Rob.* 29, 719–733. doi: 10.1109/TRO.2012.2235699

Klein, R. M. (2000). Inhibition of return. *Trends Cogn. Sci.* 4, 138–147. doi: 10.1016/S1364-6613(00)01452-2

Klink, P. C., Jentgens, P., and Lorteije, J. A. (2014). Priority maps explain the roles of value, attention, and salience in goal-oriented behavior. *J. Neurosci.* 34, 13867–13869. doi: 10.1523/JNEUROSCI.3249-14.2014

Knudsen, E. I. (2007). Fundamental components of attention. *Annu. Rev. Neurosci.* 30, 57–78. doi: 10.1146/annurev.neuro.30.051606.094256

Köster, M., Finger, H., Graetz, S., Kater, M., and Gruber, T. (2018). Theta-gamma coupling binds visual perceptual features in an associative memory task. *Sci. Rep.* 8, 1–9. doi: 10.1038/s41598-018-35812-7

Kragic, D., Björkman, M., Christensen, H. I., and Eklundh, J.-O. (2005). Vision for robotic object manipulation in domestic settings. *Rob. Auton. Syst.* 52, 85–100. doi: 10.1016/j.robot.2005.03.011

Lanillos, P. (2013). *Minimum time search of moving targets in uncertain environments* (Ph.D. thesis).

Lanillos, P., and Cheng, G. (2018b). "Adaptive robot body learning and estimation through predictive coding," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 4083–4090.

Lanillos, P., Dean-Leon, E., and Cheng, G. (2016). Yielding self-perception in robots through sensorimotor contingencies. *IEEE Trans. Cogn. Dev. Syst.* 9, 100–112. doi: 10.1109/TCDS.2016.2627820

Lanillos, P., Ferreira, J. F., and Dias, J. (2015a). "Designing an artificial attention system for social robots," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Hamburg: IEEE), 4171–4178.

Lanillos, P., Ferreira, J. F., and Dias, J. (2015b). "Multisensory 3D saliency for artificial attention systems," in *Proc. 3rd Workshop Recogn* (Action Scene Understanding), 1–6.

Lanillos, P., Gan, S. K., Besada-Portas, E., Pajares, G., and Sukkarieh, S. (2014). Multi-uav target search using decentralized gradient-based negotiation with expected observation. *Inf. Sci.* 282, 92–110. doi: 10.1016/j.ins.2014.05.054

Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., et al. (2021). Active inference in robotics and artificial agents: Survey and challenges. *arXiv[preprint].arXiv:2112.01871*. doi: 10.48550/arXiv.2112.01871

Lanillos, P., and Cheng, G. (2018a). "Active attention applications in robotics," in *International Workshop on Active Vision, Attention, and Learning* (Tokyo: IEEE Developmental Learning and Epigenetic Robotics (ICDL-Epirob)).

LaValle, S. M. (2006). *Planning Algorithms*. Cambridge: Cambridge University Press.

Lengyel, M., Yang, S. C.-H., and Wolpert, D. M. (2016). Active sensing in the categorization of visual patterns. *eLife* 5, e12215. doi: 10.7554/eLife.12215

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat*. 27, 986–1005. doi: 10.1214/aoms/1177728069

Louie, K., and Glimcher, P. W. (2019). "Normalization principles in computational neuroscience," in *Oxford Research Encyclopedia of Neuroscience*. Available online at: https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-43 (accessed July 7, 2022).

Marchant, R., and Ramos, F. (2014). "Bayesian optimisation for informative continuous path planning," in *2014 IEEE International Conference on Robotics and Automation (ICRA)* (Hong Kong: IEEE), 6136–6143.

Meera, A. A., Popović, M., Millane, A., and Siegwart, R. (2019). "Obstacle-aware adaptive informative path planning for uav-based target search," in *2019 International Conference on Robotics and Automation (ICRA)* (Montreal, QC: IEEE), 718–724.

Meera, A. A., and Wisse, M. (2020). "Free energy principle based state and input observer design for linear systems with colored noise," in *2020 American Control Conference (ACC)* (Denver, CO: IEEE), 5052–5058.

Meera, A. A., and Wisse, M. (2021a). A brain inspired learning algorithm for the perception of a quadrotor in wind. *arXiv[preprint].arXiv:2109.11971*. doi: 10.48550/arXiv.2109.11971

Meera, A. A., and Wisse, M. (2021b). "On the convergence of dem's linear parameter estimator," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (Cham: Springer International Publishing), 692–700.

Meera, A. A., and Wisse, M. (2022). Free energy principle for the noise smoothness estimation of linear systems with colored noise. *arXiv[preprint].arXiv:2204.01796*. doi: 10.48550/arXiv.2204.01796

Mirza, M. B., Adams, R. A., Friston, K., and Parr, T. (2019). Introducing a bayesian model of selective attention based on active inference. *Sci. Rep*. 9, 1–22. doi: 10.1038/s41598-019-50138-8

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci*. 10, 56. doi: 10.3389/fncom.2016.00056

Morén, J., Ude, A., Koene, A., and Cheng, G. (2008). Biologically based top-down attention modulation for humanoid interactions. *Int. J. Humanoid Rob*. 5, 3–24. doi: 10.1142/S0219843608001285

Mousavi, S., Schukat, M., Howley, E., Borji, A., and Mozayani, N. (2016). Learning to predict where to look in interactive environments using deep recurrent q-learning. *arXiv[preprint].arXiv:1612.05753*. doi: 10.48550/arXiv.1612.05753

Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). A constructive model for the development of joint attention. *Conn. Sci*. 15, 211–229. doi: 10.1080/0954009310001655101

Nakajima, M., Schmitt, L. I., and Halassa, M. M. (2019). Prefrontal cortex regulates sensory filtering through a basal ganglia-to-thalamus pathway. *Neuron* 103, 445–458. doi: 10.1016/j.neuron.2019.05.026

Nakayama, R., and Motoyoshi, I. (2019). Attention periodically binds visual features as single events depending on neural oscillations phase-locked to action. *J.Neurosci*. 39, 4153–4161. doi: 10.1523/JNEUROSCI.2494-18.2019

Oberauer, K. (2019). Working memory and attention-a conceptual analysis and review. *J. Cogn*. 2, 58. doi: 10.5334/joc.58

Oliver, G., Lanillos, P., and Cheng, G. (2021). An empirical study of active inference on a humanoid robot. *IEEE Trans. Cogn. Dev. Syst*. 14, 462–471. doi: 10.1109/TCDS.2021.3049907

Orabona, F., Metta, G., and Sandini, G. (2005). "Object-based visual attention: a model for a behaving robot," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops* (San Diego, CA: IEEE), 89–89.

Oudeyer, P.-Y., and Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Front. Neurorob*. 1, 6. doi: 10.3389/neuro.12.006.2007

Panichello, M. F., and Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature* 592, 601–605. doi: 10.1038/s41586-021-03390-w

Parr, T., Benrimoh, D. A., Vincent, P., and Friston, K. J. (2018). Precision and false perceptual inference. *Front. Integr. Neurosci*. 12, 39. doi: 10.3389/fnint.2018.00039

Parr, T., Corcoran, A. W., Friston, K. J., and Hohwy, J. (2019). Perceptual awareness and active inference. *Neurosci. Consciousness* 2019, niz012. doi: 10.1093/nc/niz012

Parr, T., and Friston, K. J. (2017a). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14, 20170376. doi: 10.1098/rsif.2017.0376

Parr, T., and Friston, K. J. (2017b). Working memory, attention, and salience in active inference. *Sci. Rep*. 7, 1–21. doi: 10.1038/s41598-017-15249-0

Parr, T., and Friston, K. J. (2019). Attention or salience? *Curr. Opin. Psychol*. 29, 1–5. doi: 10.1016/j.copsyc.2018.10.006

Parr, T., and Pezzulo, G. (2021). Understanding, explanation, and active inference. *Front. Syst. Neurosci*. 15, 772641. doi: 10.3389/fnsys.2021.772641

Parr, T., Sajid, N., Da Costa, L., Mirza, M. B., and Friston, K. J. (2021). Generative models for active vision. *Front. Neurorobot*. 15, 651432. doi: 10.3389/fnbot.2021.651432

Peters, B., Kaiser, J., Rahm, B., and Bledowski, C. (2020). Object-based attention prioritizes working memory contents at a theta rhythm. *J. Exp. Psychol. Gen*. 150, 1250–1256. doi: 10.1037/xge0000994

Phillips, J. M., Kambi, N. A., and Saalmann, Y. B. (2016). A subcortical pathway for rapid, goal-driven, attentional filtering. *Trends Neurosci*. 39, 49–51. doi: 10.1016/j.tins.2015.12.003

Pomper, U., and Ansorge, U. (2021). Theta-rhythmic oscillation of working memory performance. *Psychol. Sci*. 32, 1801–1810. doi: 10.1177/09567976211013045

Popović, M., Vidal-Calleja, T., Hitz, G., Sa, I., Siegwart, R., and Nieto, J. (2017). "Multiresolution mapping and informative path planning for uav-based terrain monitoring," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 1382–1388.

Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* 16, 1843–1848. doi: 10.1097/01.wnr.0000183900.92901.fc

Rasouli, A., Lanillos, P., Cheng, G., and Tsotsos, J. K. (2020). Attention-based active visual search for mobile robots. *Auton. Robots* 44, 131–146. doi: 10.1007/s10514-019-09882-z

Raybourn, M. S., and Keller, E. L. (1977). Colliculoreticular organization in primate oculomotor system. *J. Neurophysiol*. 40, 861–878. doi: 10.1152/jn.1977.40.4.861

Reynolds, J. H., and Heeger, D. J. (2009). The normalization model of attention. *Neuron* 61, 168–185. doi: 10.1016/j.neuron.2009.01.002

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of v4 neurons. *Neuron* 26, 703–714. doi: 10.1016/S0896-6273(00)81206-4

Rizzolatti, G., Riggio, L., Dascola, I., and Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia* 25, 31–40. doi: 10.1016/0028-3932(87)90041-8

Roberts, R., Ta, D.-N., Straub, J., Ok, K., and Dellaert, F. (2012). "Saliency detection and model-based tracking: a two part vision system for small robot navigation in forested environment," in *Unmanned Systems Technology XIV, Vol. 8387* (International Society for Optics and Photonics), 83870S.

Rucci, M., Ahissar, E., and Burr, D. (2018). Temporal coding of visual space. *Trends Cogn. Sci*. 22, 883–895. doi: 10.1016/j.tics.2018.07.009

Ruff, D. A., and Cohen, M. R. (2016). Stimulus dependence of correlated variability across cortical areas. *J. Neurosci*. 36, 7546–7556. doi: 10.1523/JNEUROSCI.0504-16.2016

Sajid, N., Da Costa, L., Parr, T., and Friston, K. (2021a). Active inference, bayesian optimal design, and expected utility. *arXiv[preprint].arXiv:2110.04074*. doi: 10.1017/9781009026949.007

Sajid, N., Faccio, F., Da Costa, L., Parr, T., Schmidhuber, J., and Friston, K. (2021b). Bayesian brains and the r\'enyi divergence. *arXiv[preprint].arXiv:2107.05438*. doi: 10.48550/arXiv.2107.05438

Sajid, N., Friston, K. J., Ekert, J. O., Price, C. J., and Green, D. W. (2020). Neuromodulatory control and language recovery in bilingual aphasia: An active inference approach. *Behav. Sci*. 10, 161. doi: 10.3390/bs10100161

Sajid, N., Holmes, E., Costa, L. D., Price, C., and Friston, K. (2022). A mixed generative model of auditory word repetition. *bioRxiv [preprint]*. doi: 10.1101/2022.01.20.477138

Santangelo, V. (2015). Forced to remember: when memory is biased by salient information. *Behav. Brain Res*. 283, 1–10. doi: 10.1016/j.bbr.2015.01.013

Santangelo, V., Di Francesco, S. A., Mastroberardino, S., and Macaluso, E. (2015). Parietal cortex integrates contextual and saliency signals during the encoding of natural scenes in working memory. *Hum. Brain Mapp*. 36, 5003–5017. doi: 10.1002/hbm.22984

Santangelo, V., and Macaluso, E. (2013). Visual salience improves spatial working memory via enhanced parieto-temporal functional connectivity. *J. Neurosci*. 33, 4110–4117. doi: 10.1523/JNEUROSCI.4138-12.2013

Schmitz, T. W., and Duncan, J. (2018). Normalization and the cholinergic microcircuit: a unified basis for attention. *Trends Cogn. Sci*. 22, 422–437. doi: 10.1016/j.tics.2018.02.011

Shon, A. P., Grimes, D. B., Baker, C. L., Hoffman, M. W., Zhou, S., and Rao, R. P. (2005). "Probabilistic gaze imitation and saliency learning in a robotic head," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (Barcelona: IEEE), 2865–2870.

Sommer, M. A., and Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature* 444, 374–377. doi: 10.1038/nature05279

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vis. Res*. 48, 1391–1408. doi: 10.1016/j.visres.2008.03.009

Tomassini, A., Ambrogioni, L., Medendorp, W. P., and Maris, E. (2017). Theta oscillations locked to intended actions rhythmically modulate perception. *Elife* 6, e25618. doi: 10.7554/eLife.25618

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol*. 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artif. Intell*. 78, 507–545. doi: 10.1016/0004-3702(95)00025-9

Ude, A., Wyart, V., Lin, L.-H., and Cheng, G. (2005). "Distributed visual attention on a humanoid robot," in *5th IEEE-RAS International Conference on Humanoid Robots, 2005* (Tsukuba: IEEE), 381–386.

VanRullen, R. (2016). Perceptual cycles. *Trends Cogn. Sci*. 20, 723–735. doi: 10.1016/j.tics.2016.07.006

Welch, G., and Bishop, G. (2002). *An Introduction to the Kalman Filter*. Chapel Hill, NC: University of North Carolina.

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., and Munoz, D. P. (2017). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nat. Commun*. 8, 1–9. doi: 10.1038/ncomms14263

Whiteley, L., and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *J. Vis*. 8, 2–2. doi: 10.1167/8.3.2

Yang, S. C.-H., Lengyel, M., and Wolpert, D. M. (2016a). Active sensing in the categorization of visual patterns. *Elife* 5, e12215. doi: 10.7554/eLife.12215

Yang, S. C.-H., Wolpert, D. M., and Lengyel, M. (2016b). Theoretical perspectives on active sensing. *Curr. Opin. Behav. Sci*. 11, 100–108. doi: 10.1016/j.cobeha.2016.06.009

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fnbot.2022.910161

Check for updates

OPEN ACCESS

EDITED BY
Inês Hipólito,
Humboldt University of Berlin, Germany

REVIEWED BY
Andrei Khrennikov,
Linnaeus University, Sweden
Wen-Ran Zhang,
Georgia Southern University, United States
Aswani Kumar Cherukuri,
VIT University, India

*CORRESPONDENCE
Yukio-Pegio Gunji
yukio@waseda.jp

RECEIVED 01 April 2022
ACCEPTED 07 July 2022
PUBLISHED 02 September 2022

CITATION
Gunji Y-P, Shinohara S and Basios V (2022) Connecting the free energy principle with quantum cognition. *Front. Neurorobot.* 16:910161. doi: 10.3389/fnbot.2022.910161

COPYRIGHT
© 2022 Gunji, Shinohara and Basios. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Connecting the free energy principle with quantum cognition

Yukio-Pegio Gunji[1]*, Shuji Shinohara[2] and Vasileios Basios[3]

[1]Department of Intermedia Arts and Science, School of Fundamental Science and Technology, Waseda University, Tokyo, Japan, [2]School of Science and Engineering, Tokyo Denki University, Tokyo, Japan, [3]Service de Physique des Systèmes Complexes et Mécanique Statistique and Interdisciplinary Center for Nonlinear Phenomena and Complex Systems (C.P.231 CeNoLi-ULB), Université Libre de Bruxelles, Brussels, Belgium

It appears that the free energy minimization principle conflicts with quantum cognition since the former adheres to a restricted view based on experience while the latter allows deviations from such a restricted view. While free energy minimization, which incorporates Bayesian inference, leads to a Boolean lattice of propositions (classical logic), quantum cognition, which seems to be very dissimilar to Bayesian inference, leads to an orthomodular lattice of propositions (quantum logic). Thus, we address this challenging issue to bridge and connect the free energy minimization principle with the theory of quantum cognition. In this work, we introduce "excess Bayesian inference" and show that this excess Bayesian inference entails an underlying orthomodular lattice, while classic Bayesian inference entails a Boolean lattice. Excess Bayesian inference is implemented by extending the key idea of Bayesian inference beyond classic Bayesian inference and its variations. It is constructed by enhancing the idea of active inference and/or embodied intelligence. The appropriate lattice structure of its logic is obtained from a binary relation transformed from a distribution of the joint probabilities of data and hypotheses by employing a rough-set lattice technique in accordance with quantum cognition logic.

KEYWORDS

**free energy minimization, quantum cognition, Bayesian inference, rough set, lattice theory**

## Introduction

Cognitive predictive behaviors that are found in brain function, biological information processing, and cognitive sciences have been recently described and explained using the free energy minimization principle (Friston et al., 2006; Friston and Kiebel, 2009a,b). However, related cognitive phenomena such as sensory illusions (e.g., due to ambiguity such as in the Necker cube), the conjunction fallacy (e.g., "Linda's fallacy"), the order effect in questionnaire responses, context-dependent decision-making, and the "Guppy effect" in complex concept conjunction and disjunction have been recently described and explained by using quantum cognition principles (Khrennikov, 2001, 2010, 2021; Aerts, 2009; Aerts et al., 2012, 2013, 2019; Busemeyer and Bruza, 2012; Haven and Khrennikov, 2013; Asano et al., 2015; Bruza et al., 2015; Dzhafarov et al., 2016; Ishwarya and Kumar, 2020a,b). Moreover, recent developments

in quantum measurement theory provide a general mathematical framework that can accommodate the question order effect and the response replicability effect as well as their combinations. Thus, the generalization of the Wang–Busemeyer quantum-cognition postulates for quantum-like modeling of decision-making and psychology is achieved. An up-to-date discussion of these recent developments and an introduction to this "theory of quantum instruments" can be found in Ozawa and Khrennikov (2021).

It has come to our attention through a reviewer's suggestion that in a recent publication (Zhang, 2021a,b), a newly proposed analytical quantum computing paradigm, called "quantum intelligence" (QI), aims at elucidating the notion of causality concerning the underlying logic of the phenomena under scrutiny.

Predictive algorithms and coding that deal with large problems, i.e., problems that require too much computation to obtain an optimal solution, meet the seriously challenging problem of reducing the search area for the solutions to make their implementation manageable or even possible. Bayesian inference is one of the most powerful techniques that can solve this problem (Arecchi, 2003, 2011). Moreover, it is well known that Bayesian inference can be accurately formulated as an instance, or a case of the free energy minimization principle (Friston et al., 2006; Friston, 2010). The efficacy of Bayesian inference is because it only focuses on an a priori probability distribution assumed by the already given or realized events. These events have been "experienced," "realized," or "recorded," as given, but there is no requirement that *all* events from such an a priori distribution have been "realized." The relevant events could all, in principle, be hypothetical. In the context of the free energy principle, the prior distribution is implemented by a generative model of the action of the environment on a Markov blanket. As this action is generated outside the Markov blanket, by the dynamics of the environment, it is unobservable "in principle," i.e., under this prior. Therefore, such actions are duly ignored; obviously, this inherent aspect of Bayesian inference helps to reduce the search space. This is the essence of Bayesian inference.

Furthermore, an embodied mind and/or intelligence (Varela et al., 1991; Varela, 1997) can augment and complement the process of Bayesian inference in biological information processing and even overcome some of its drawbacks (Seth and Friston, 2016; Allen and Friston, 2018; Seth and Tsakiris, 2018; Yon et al., 2019; Walsh et al., 2020). For example, let us assume that you are familiar only with the front face of your preferred singer, say through some photos. Now imagine that you encounter that singer on a street in a town; then, you would naturally be impelled to move your body to try to see the front face. Since you have experienced more front-face photos than photos from other angles, Bayesian inference would ignore the face data that have been experienced less often, and this would make the relation between the singer in the street and the familiar front-face image retrieved by memory stronger; eventually, this would result in the recognition that this person is indeed your preferred singer, provided that the match is close enough. This is a result of the fact that you cannot identify the face seen from an angle (because of uncertainty, missing data, or the frame problem), which implies that there is a disadvantage to Bayesian inference. Moving your body to see the front face of the singer implies that action-motion control and determining the correct placement of the body can address this disadvantage of Bayesian inference. This is called active or embodied inference. Therefore, since embodiment complements and reinforces Bayesian inference, one can state that a stubbornly predictive coding is stably generated in the brain. These ideas are also implemented in robotics (Linson et al., 2018; Çatal et al., 2021).

While Bayesian inference seems to be flexible and to be far from rigorous logical thinking, it entails nothing but classical logic, or in other words, Boolean algebra, in which any phenomenon can be explained by a combination of atomic propositions taking the values yes or no and subject to the law of excluded middle (i.e., classical logical reductionism). Bayesian inference itself is not flexible in its logic, but it is a flexible *method* to determine within a predictive area what can be assessed and accepted by classical Boolean algebra (Arecchi, 2003, 2011; Gunji et al., 2017, 2020). In that sense, an image encoded inside the brain and an object existing outside of the brain must have a one-to-one correspondence through predictive coding.

In contrast, quantum cognition focuses on the other side of brain-function phenomena and/or cognition (Khrennikov, 2001, 2021; Aerts, 2009; Aerts et al., 2012, 2013, 2019; Haven and Khrennikov, 2013; Asano et al., 2015; Bruza et al., 2015). Quantum cognition describes and explains cognition, apprehension, comprehension, perception, and decision-making by using the basic formalism and conceptual logical and mathematical framework of quantum mechanics. It is not concerned with the physical basis of quantum processes in the brain's microscopic dynamics, and it does not apply quantum mechanics to macroscopic phenomena such as cognition and perception. This is why the quantum cognition community argues in favor of quantum mechanics being properly utilized only as a mathematical tool to model cognitive phenomena. Quantum logic, on which quantum cognition is based, fundamentally differs from Boolean logic (Boolean algebra). While Boolean logic has the structure of a rather simple complemented distributive lattice, quantum logic has the structure of a non-distributive lattice, i.e., a more complicated orthomodular lattice. This more complex structure implies that the cognition of multiple events entails a kind of resonance. This means that these multiple events interfere with each other due to the non-distributive nature of their logical evaluation, resulting in mutual or multi-interdependence. Resonance typically implies a non-linear interdependence (here in evaluating probabilities); distributive lattices have linear dependencies for

their propositions and hence no resonance-type effects. The distributive law guarantees the independence of its reduced atomic events or, in other words, atomic propositions. In contrast, a non-distributive lattice guarantees the emergence of interactions among reduced atomic events, which entails resonance-like aspects for the probability of independent events. This can explain cognitive illusions such as the conjunction fallacy, in which the joint probability of the occurrence of events *A* and *B* is larger than that of the occurrence of *A*.

Quantum logic (i.e., an orthomodular lattice) results from a property of Hilbert space for the operators in quantum mechanics. While there is no fundamental physical reason to assume Hilbert space in a macroscopic universe, classical mechanics can still be formulated in a Hilbert space framework, as Koopman and Von Neumann proposed in the 1930s. However, this results in an operational probabilistic theory endowed with a classical Boolean algebra, which is complete as a lattice. The dependence on the Hilbert space framework for quantum cognition was considered a problem, or at least an inconvenience, in assigning meaning to the related operators and spaces. Quantum logic has been exemplified by Dirac's famous 3-polarizer experiment, but recently (Zhang, 2021a), an analysis for bipolar crisp and fuzzy sets has provided new insights into the old question. Indeed, as is well known, fuzzy set theory models vagueness by membership measures, while rough sets model incomplete information by bounding it with a lower and an upper approximation. Therefore, it would be very interesting to consider our previous work on rough set approximation and quantum cognition (Gunji and Haruna, 2010; Gunji et al., 2016; Gunji and Nakamura, 2022a,b) in this new light in future investigations. This could prove instrumental in taking further steps toward a deeper understanding of the interrelation of Bayesian inference and causal inference.

However, it has been recently verified that quantum logic, or orthomodular lattices, can be constructed without Hilbert space. First, this was achieved (Gunji et al., 2016) by the extension of Arecchi's idea of inverse Bayesian inference (Arecchi, 2003, 2011), and then it was achieved by the idea of ambiguity between what is inside and outside of a context (Gunji and Haruna, 2022; Gunji and Nakamura, 2022a,b). It can be achieved with respect to "rough set" lattices—a kind of special coarse-graining operation on regular sets—based on a binary relation. Quantum logic without Hilbert space has also been achieved by using category theory (Heunen and Vicary, 2019). This implies that there is now a clear and reasonable foundation by which quantum logic structures can readily be applied to macroscopic phenomena. Now, we can turn to the following questions that arise:

How are Bayesian inference and quantum cognition interrelated in macroscopic world phenomena?

While classic Bayesian inference leads to Boolean logic, in which classical logical reductionism holds, quantum logic can never be compatible with classical logical reductionism. Does this imply that the coexistence of Bayesian inference and quantum logic in a macroscopic setting entails an antinomy? How is Bayesian inference, which can be cast in a free-energy-principle form, interrelated with quantum logic or orthomodular lattices? If a datum is not related to the prior, then there is no context or its probability is recorded as zero. In other words, the probability of an event outside the context is almost zero. Thus, the underlying logic leads to a Boolean lattice. In contrast, quantum logic (or an orthomodular lattice) allows for contextuality since it accepts the non-zero probability of an event even outside the originally set context. We show here that such a non-zero probability of an event outside the original context can be obtained from an *"excessive Bayesian procedure"* or, in other words, an *"extended Bayesian inference"*. This entails a variation of the so-called *"Bayesian-Inverse-Bayesian"* non-linear loop (Gunji et al., 2016; Basios and Gunji, 2021). It might seem paradoxical, but upon closer examination, it is not. Although a one-to-one correspondence is enforced within the context, the non-zero probability of an event outside the given context can still be readily obtained. Stubborn predictive coding is resistant to change, and seemingly paradoxically, it not only affords but also actually gives rise to the possibility of considering other "outsider" events in addition to the events inside the context. This results in an instance of quantum logic.

This article is organized as follows: First, we show the relationship between the free energy minimization principles—Bayesian inference and Boolean algebra. Second, we implement an excess Bayesian procedure and demonstrate how the relationship between the datum and hypothesis is changed through this procedure. Third, we show that the excess Bayesian procedure entails an orthomodular lattice as a quasi-disjoint union of Boolean algebras. Therefore, in conclusion, we establish that this implies that quantum cognition and the free energy principle are connected to each other *via* an excess Bayesian procedure.

## Quantum cognition, orthomodular lattice (quantum logic), and free energy minimization

Quantum cognition, which is a new trend in cognitive science, is based on the notion of probability in quantum mechanics. Since any state of an event is defined as a vector of complex numbers, the probability of an event is expressed as the norm of the vector, as in quantum mechanics. Since the effect of quantum entanglement plays an essential role in calculating the joint probability, quantum cognition can explain various cognitive illusions. However, quantum cognition uses quantum mechanics not as a physical foundation of cognition but as information theory.

The orthomodular lattice is directly obtained from quantum mechanics; a lattice is an ordered set that is closed with respect to

binary operations, *meet* and *join* (Appendix A). Given a complex number linear space, an element of an ordered set is defined by a set of vectors, and the order relation is defined by inclusion. For any set of vectors, a set of vectors that are orthogonal to the vectors is defined as an orthocomplement of them. The *meet* of sets of vectors is defined by their intersection, and the *join* of sets is defined by the composition of the orthocomplement of their union. Thus, an ordered set can be verified as an orthocomplemented lattice. Since a linear vector space is equipped with a Hilbert space, an orthocomplemented lattice is verified as an orthomodular lattice. If inclusion is regarded as a sequence of premises and consequences and *meet* and *join* are regarded as logical conjunction and disjunction, respectively, an orthomodular lattice is reformalized as quantum logic.

While quantum cognition and the orthomodular lattice were originally derived from quantum mechanics equipped with a Hilbert space, it has been verified that an orthomodular lattice can be obtained without a Hilbert space, and various cognitive illusions can be explained by subjective probability defined in an orthomodular lattice without a Hilbert space. This implies that quantum cognition may be established by an orthomodular lattice alone, without a Hilbert space.

Free energy minimization is a theory by which cognitive brain function is systematically explained, and it is based on Bayesian inference. The term free energy originates from the fact that the upper bound of the cost function that reveals the difficulty in predicting the sensory input is called variational free energy. First, we spell out how free energy minimization is related to Bayesian inference and how it is disconnected from quantum cognition. It is well known that the free energy minimization principle implies Kullback–Leibler divergence between the a priori probability and the a posteriori probability under the minimization of predictive error (Friston and Kiebel, 2009a,b). It is expressed as

$$min\left(KLD\left[p\left(u\right)\parallel p\left(u|s\right)\right]-\ln p\left(s\right)\right) \tag{1}$$

Where, $KLD\left[p\left(x\right)\parallel q\left(x\right)\right]=\sum_i p\left(x_i\right)\log\frac{p(x_i)}{q(x_i)}$ is the Kullback–Leibler divergence between the two given probability distributions, $p$, $q$, and the random variable is $x$, while $s$ is the given datum that has been experienced, and $lnp\left(s\right)$ is the surprise resulting from a given experience (i.e., the predictive error), which is minimized. It is easy to see that $KLD\left[p\left(x\right)\parallel q\left(x\right)\right]=0$ if, and only if, $p\left(x\right)=q\left(x\right)$ almost everywhere.

Thus, the minimizing procedure (1) implies that the a priori probability coincides with the a posteriori probability. Here, we show that this procedure is nothing other than classic Bayesian inference. Since probability changes over time in Bayesian inference, let us introduce time as a suffix for the probability. The variables *d and h* represent the datum and hypothesis, respectively. The probability of datum $d$ at time step $t$ is represented by $P^t\left(d\right)$, and that of hypothesis $h$ at time step $t$ is represented by $P^t\left(h\right)$.

The conditional probability $P^t\left(h|d\right)$ represents the probability of $h$ under the experience of $d$. Since hypothesis $h$ is the probability distribution of the data, hypothesis $h$ is expressed in terms of the likelihood of data as $P^t\left(d|h\right)$.

From the definition of the conditional probability of $A$ given $B$, expressed as $P\left(A|B\right)$, we have $P\left(d\middle|h\right)=\frac{P(d,h)}{P(h)}$ and $P\left(h\middle|d\right)=\frac{P(d,h)}{P(d)}$, so one obtains *via* Bayes' theorem that

$$P\left(h\middle|d\right)P\left(d\right)=P\left(d\middle|h\right)P\left(h\right) \tag{2}$$

This is consistent with non-Bayesian probability theory when there is no iteration over time. However, because $P\left(d\right)=\sum_k P\left(d|h_k\right)P\left(h_k\right)$, and because we do have an iterative procedure over time in our case, with a time step $t$, we obtain for $P^t\left(h\right)$ the following expression:

$$P^t\left(h\middle|d\right)=\frac{P^t\left(d|h\right)P^t\left(h\right)}{\sum_k P^t\left(d|h_k\right)P^t\left(h_k\right)} \tag{3}$$

One might regard the calculation using Equation 3 as Bayesian inference, yet this is not a genuine Bayesian inference just because it is consistent with ordinary probability theory. The essence of Bayesian inference is that it allows us to compute, starting from a given a priori probability, a resulting a posteriori probability such that

$$P^{t+1}\left(h\right)=P^t\left(h|d\right) \tag{4}$$

This is the goal of free energy minimization. The probability of the hypothesis $h$ under a specific experience $d$ is generalized as the probability of the hypothesis independent of experience.

The relation between data and hypotheses is analogous to the relationship between objects outside the brain and their representations (or "images") inside the brain. Let us consider a set of hypotheses and data, $H=\left\{h_1,h_2,\ldots,h_N\right\},D=\left\{d_1,d_2,\ldots,d_N\right\}$. A one-to-one correspondence between hypotheses and data is expressed by using the likelihood of a hypothesis and the probability of this hypothesis. The likelihood is expressed as

$$P^t\left(d_i\middle|h_j\right)\sim 1.0,\left(i=j\right);P^t\left(d_i\middle|h_j\right)\sim 0.0,\left(i\neq j\right). \tag{5}$$

This one-to-one correspondence between hypotheses and data implies $P^t\left(h\right)_j=\frac{1}{N}$.

Therefore, the joint probability of a hypothesis is expressed as

$$P^t\left(d_i,h_j\right)=P^t\left(d_i\middle|h_j\right)P^t\left(h_j\right)\sim\frac{1}{N},\left(i=j\right);P^t\left(d_i,h_j\right)\sim 0.0,\left(i\neq j\right). \tag{6}$$

The joint probability between a hypothesis and data is transformed into a binary relation $R\subseteq H\times D$ such that $\left(h,\ d\right)\in R$ if $P^t\left(d,\ h\right)>\theta$; otherwise, $\left(h,\ d\right)\notin R$,.

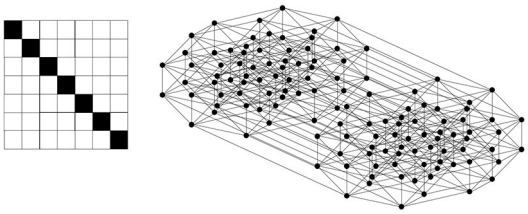


FIGURE 1
A diagonal relation and its corresponding lattice, which is a Boolean lattice. The lattice is shown here as a Hasse diagram, in which the elements of the lattice (a subset of $D$) are represented by black circles, and if one element is smaller than another element (i.e., one set is included in the other) and no element exists between them, then they are connected by a line, and the larger one is shown above the smaller one.

This results in a diagonal relation, as shown in Figure 1. This construction of a binary relation can be generalized to any joint probability between a hypothesis and data. In Figure 1, $H = \{h_1, h_2, \ldots, h_7\}, D = \{d_1, d_2, \ldots, d_7\}$, and from Equation 6, $P^t(d_i, h_i) \sim 0.143$; $P^t(d_i, h_j) \sim 0.0$. Thus, given $\theta = 0.1$, we obtain $(d_i, h_i) \in R \subseteq D \times H$ for $i = 1, 2, \ldots, 7$ and $(d_i, h_j) \notin R$ for $i \neq j$.

Now, given a binary relation between a hypothesis and data, one can estimate a logical structure with respect to a lattice (see Appendix A). There are several ways to construct a lattice from a binary relation, where a *concept* is defined as a fixed point with respect to certain defined operations. This has the result that a *concept* is expressed as a pair of a subset of the hypotheses, $H$, and a subset of the data, $D$. Since such a *concept* formulation for a rough set lattice is consistent with the "concept" discussed in cognitive linguistics, we apply the rough set lattice formalism and theory for a binary relation.

Given a set of hypotheses $H$, a set of data $D$, and a relation $R \subseteq H \times D$, two operations, the upper and lower approximations (Appendix B) are defined as follows. For $\forall X \subseteq D$, the upper approximation of $X$ with respect to a hypothesis is defined by

$$H^*(X) = \{T \in H | s \in X, sRT\} \tag{7}$$

For $\forall Y \subseteq H$, the lower approximation of $Y$ with respect to the data is defined by

$$D_*(Y) = D - \{s \in D | T \in H - Y, sRT\} \tag{8}$$

where $H - Y$ represents the complement of $Y$ in the set $H$.

A collection of fixed points of the composition of the two operations (7) and (8) is called a rough set lattice (Yao, 2004; Gunji and Haruna, 2010; see Appendix-B) and is described as

$$L = \left\{X \subseteq D \middle| D_*\left(H^*(X)\right) = X\right\} \tag{9}$$

As shown in Appendix A, a lattice is defined by an ordered set that is closed with respect to specific binary operations, *meet* and *join*. In a rough set lattice, $L$ is a subset of the power set of $D$ and is closed with respect to *join* and *meet*. While there are other methods for constructing a lattice from a binary relation, such as the lattice of formal concept analysis that was developed to deal with cognitive memory (Kumar et al., 2015; Shivahare and Cherukuri, 2017), we used a rough set lattice for reasons based on cognitive linguistics.

In the ideal case of a diagonal relation, as shown in Figure 1, one obtains a Boolean lattice (Appendix A) of a rough-set lattice. First, $D_*\left(H^*\left(\{d_i\}\right)\right) = D_*\left(\{h_i\}\right) = \{d_i\}$ for any $i = 1, 2, \ldots, 7$. Thus, a singleton set of any element of $D$ satisfies Equation 9 and is an element of a rough set lattice. Since a Boolean lattice is expressed as a power set of $D$, any subset of $D$ is an element of $L$. This is easy to verify since for any subset of $D$, such as $X = \{d_i, d_j, \ldots\}$,

$$D_*\left(H^*(X)\right) = D_*\left(\{h_i, h_j, \ldots\}\right) \tag{10}$$

and since for any $k \in \{1, 2, \ldots, N\}, (h_k, d_k) \in R$, and $(h_s, d_k) \notin R, (s \neq k)$. Indeed, for any subset of $D$,

$$\begin{aligned} D_*\left(\{h_i, h_j, \ldots\}\right) &= D - \left\{s \in D | T \in \overline{\{h_i, h_j, \ldots\}}, sRT\right\} \\ &= D - \overline{\{d_i, d_j, \ldots\}} = \{d_i, d_j, \ldots\} \end{aligned} \tag{11}$$

where $\overline{\{d_i, d_j, \ldots\}}$ is the complement of $\{d_i, d_j, \ldots\}$.

This implies that for any $X \subseteq D$,

$$D_*\left(H^*(X)\right) = X \tag{12}$$

Therefore, $L$ is the same as the power set of $D$, and *meet* is defined by intersection while *join* is defined by union. A Boolean lattice is a classical set-theoretic logic, and any *concept* within this logic, as defined above, can be expressed as a combination of logical atoms (i.e., the next least element of $D$), which in turn implies that any such *concept* can be reduced to atoms. This is why the Boolean lattice is simply classical logical reductionism. In the Hasse diagram of Figure 1, all elements of the rough set lattice defined by the power set of $D = \{d_1, d_2, \ldots, d_7\}$ are represented by black circles. The least element is the empty set, and the elements, called atoms, just above the least element are $\{d_1\}, \{d_2\}, \ldots, \{d_7\}$. Then, the elements just above the atoms are all combinations of atoms, such as $\{d_1, d_2\}, \{d_1, d_3\}, \ldots, \{d_2, d_3\}, \ldots, \{d_6, d_7\}$. The lattice contains all subsets of $D$, and the top element (i.e., the greatest element) is $D$.

A Boolean lattice is mathematically defined as a distributive complemented lattice. *Meet* and *join* constitute distributive laws for any element of the Boolean lattice. A *complemented* lattice implies that for any element there is at least one complement of it such that the *meet* of the element and its complement is the least element of the lattice, and their *join* is the greatest element of the lattice. In a Boolean lattice, for any element of the lattice,

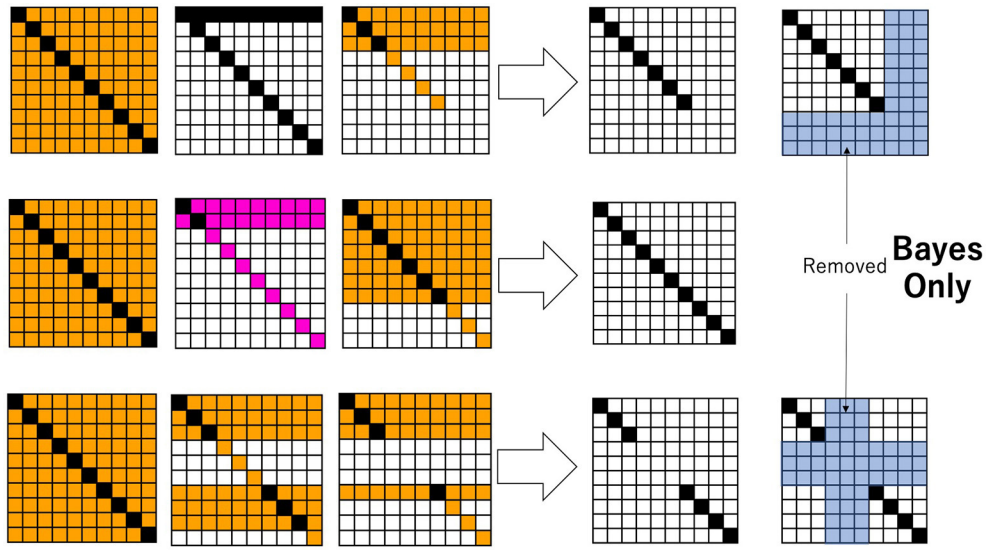



FIGURE 2
Snapshots of the joint probability between data and hypotheses. The probabilities are colored from low to high in the order of white, pale yellow, orange, pink, and black. By following the thick arrow, the binary relation $R$ is obtained. The pale blue cells represent the domain in the relation, which is ignored by Bayesian inference.

there is a unique complement called the *orthocomplement* (see Appendix A).

Now, instead of an ideal diagonal relation, let us consider a binary relation between hypothesis and datum through Bayesian inference. Given an ideal diagonal relation as an initial condition, a specific sequence of data, $d \in D_E \subseteq D$, is given to obtain a decision-making system based on Bayesian inference. Figure 2 shows some snapshots, in the form of a heatmap, of the joint probabilities of hypotheses and data. In Figure 2, $H = \{h_1, h_2, \ldots, h_{10}\}, D = \{d_1, d_2, \ldots, d_{10}\}$, and initially, $P^{t=0}(d_i|h_i) = 0.8$, and in the case of $i \neq j$, $P^{t=0}(d_j|h_i) = \frac{1-0.8}{9} = 0.022$. The probability of each hypothesis is such that $P^{t=0}(h) = 0.1$. The temporal development follows Equations 3, 4, where specific data, $d$ in $P^t(h|d)$, are given at each time. In the top row in Figure 2, a specific $d$ is randomly given from a subset of $D$ such as $D - \{d_8, d_9, d_{10}\} = \{d_1, d_2, \ldots, d_7\}$. Since the probability of $d$ is calculated cumulatively, as time proceeds, the probability converges to the actual situation: $P^t(d_8) = P^t(d_9) = P^t(d_{10}) = 0.0$, and $P^t(d_s) = 1/7 = 0.14$ with $s \neq 8, 9, 10$. If at time step $t$, $d_s$with $s \neq 8, 9, 10$ is given, $P^t(h|d_s) = \frac{P^t(d_s|h)P^t(h)}{\sum_k P^t(d_s|h_k)P^t(h_k)}$ is calculated for any $h$. From this, $P^{t+1}(h) = P^t(h|d_s)$ is obtained by Equation 9. Finally, $P^{t+1}(d, h)$ is calculated by $P^{t+1}(d)P^{t+1}(h)$. Each matrix of Figure 2 is obtained as a heatmap, in which if $P^{t+1}(d, h) \geq 0.01$, the cell is painted black; if $0.01 > P^{t+1}(d, h) \geq 0.008$, it is painted pink; if $0.008 > P^{t+1}(d, h) \geq 0.002$, it is painted orange; if $0.002 > P^{t+1}(d, h) \geq 0.0006$, it is painted pale yellow; and otherwise, it is painted white.

It is easy to see that data that have not been experienced or realized and their corresponding hypotheses are ignored through Bayesian inference. Therefore, a binary relation $R$ is readily obtained for any $d_i \in D_E, (h_j, d_i) \in R, (i = j); (h_j, d_i) \notin R, (i \neq j)$, as shown in the right row of Figure 2. Henceforth, we call any such relation a sub-diagonal relation since a full diagonal relation holds only for a subset of $D$.

A rough set lattice corresponding to a sub-diagonal relation is also a Boolean lattice, the same as the lattice shown in Figure 1. It is easy to see that for any $X \subseteq D_E$, $D_*(H^*(X)) = X$, and for any $Y \subseteq \overline{D_E}$, $D_*(H^*(Y)) = D_*(\varnothing) = \varnothing$. Therefore, $D_*(H^*(X \cup Y)) \neq X \cup Y$, which implies that a rough-set lattice for a sub-diagonal relation on $D_E$ is the power set of $D_E$. Obviously, this is nothing more than a Boolean lattice by itself. The diagonal relation that entails a Boolean lattice also implies a one-to-one correspondence between objects and representations (or "images"), which is actually the basis for decision-making based on classical logical reductionism. When a decision-maker searches for an optimal solution through Bayesian inference, the domain in which the one-to-one correspondence holds is restricted to a small part of a whole binary relation. That small area results in a sub-diagonal relation, which helps the decision-makers avoid redundant searches. In other words, Bayesian inference is essentially the core of "stubborn empiricism." While many researchers claim that predictive coding in the brain is flexible and plastic, with respect to search strategies in a changeable environment, Bayesian inference itself sticks to a "stubborn" optimization process.

## Excess Bayesian inference and quantum logic

Compared to Bayesian inference, or its equivalent free energy minimization, quantum cognition claims remarkably better flexibility and deviates from classical optimization in terms of decision-making processes. One of the most intriguing examples is the so-called "guppy effect," which shows the essence of the conjunctive fallacy: given two events that are independent of each other, the joint probability of the two events occurring simultaneously is smaller than the probability of every single event (Aerts et al., 2012). The probability of an unknown person being male, represented by $P\left(male\right)$, is $\frac{1}{2}$, and the probability that the person is born in March, $P\left(March\right)$, is $\frac{1}{12}$. Therefore, since the probability that the person is male and is born in March is $P\left(male \wedge March\right) = \frac{1}{24}$, we have

$$P\left(male \wedge March\right) < P\left(male\right), P\left(male \wedge March\right) < P\left(March\right) \tag{13}$$

The guppy effect contradicts this situation. If someone is asked to give an example of a fish, most likely, the person will respond with tuna, mackerel, or a similar example, and the probability that the person will give the example of a guppy, the popular pet fish that gave its name to this effect in the original publication, is very small. The probability of recalling the datum "guppy" when the prompt "fish" is given is represented by $P_{guppy}\left(fish\right)$. Analogously, if someone is asked to give an example of a pet, the person might respond with a cat or a dog, and probability of recalling $P_{guppy}\left(pet\right)$ is very small. However, if specifically asked to recall a pet fish, the person could recall the guppy with high probability, and this implies

$$P\left(pet \wedge fish\right) > P\left(fish\right), P\left(pet \wedge fish\right) > P\left(pet\right) \tag{14}$$

This is the guppy effect.

In quantum cognition, quantum mechanics is used as the basis of information theory. An event is defined as a vector in Hilbert space, and the probability of an event is defined by the square of the norm of its vector. The guppy effect, in which a joint probability is larger than the probability of a single event, can be explained by the *entanglement* of events. Thus, we arrive at the realization that the probabilities of events are not independent of each other and are in fact an interaction among events. Their *"entanglement"* can occur through an effect outside the diagonal relation that results from classic Bayesian inference.

The difference between Bayesian inference and quantum cognition is clearly shown in terms of lattice theory. As mentioned previously, Bayesian inference entails a Boolean lattice. In quantum mechanics, one can define a set of vectors, $X$, as an element of a lattice, and an orthocomplement of $X$, represented by $X^{\perp}$, is defined by the set of vectors whose inner product with any vector of $X$ is zero (i.e., the two vectors are orthogonal to each other). *Meet*, again, is defined here by the intersection, and *join* is defined by the orthocomplement of the union. In Hilbert space, the resulting lattice is known to be an orthomodular lattice (Appendix A), which is essentially what quantum logic is. As mentioned previously, it has been recently demonstrated that quantum logic (an orthomodular lattice) can be constructed without a Hilbert space using category theory and/or lattice theory.

Since an orthomodular lattice is a non-distributive lattice, not all events and phenomena can be explained by a combination of logical atoms. In other words, the orthomodular lattice conflicts with the classical logical reductionistic approach supported by classical Boolean lattice theory. It can indeed accommodate flexible and plastic inference processes and interactions of thoughts concerning multiple events. In that sense, an orthomodular lattice or quantum logic not only conflicts with Boolean lattice theory but also surpasses it.

Thus far, we have obtained a lattice from a binary relation between data and hypotheses, and we have shown that Bayesian inference entails a Boolean lattice. Moreover, as has been clearly seen, Bayesian inference plays an essential role in human decision-making. Therefore, any other decision-making process that should be added to the Bayesian inference process must preserve and accommodate Bayesian inference itself. However, since decision-making based on quantum logic conflicts with decision-making based on Bayesian inference, these two decision-making processes might seem to be in conflict with each other. Therefore, the problem that arises is to determine what kind of process entailing quantum logic could be added to the basic Bayesian process.

Here, we define the *excess Bayesian process*, by which the tendency to stick to experience based on Bayesian inference is enhanced rather than canceled. The excess Bayesian process is defined by the detection of data that have been experienced and actually contracts the universe of discourse based on experience. Since the probability of a specific datum that is not observed decreases to almost 0.0 through Bayesian inference, only the joint probabilities of specific pairs of datums and hypotheses remain, which is called the effective domain. The excess Bayesian process regards the effective domain as a set consisting of all joint probabilities and modifies the likelihood function of the hypothesis so that it is contained in the effective domain. Therefore, the joint probability in the effective domain is divided by the summation of those joint probabilities in the effective domain. This is defined below, and the algorithmic representation is shown in Appendix C. While excess Bayesian inference has sometimes been referred to as inverse Bayesian inference (Gunji et al., 2016), inverse Bayesian inference is formalized so that it is symmetrical to Bayesian inference (Gunji et al., 2018, 2021). Since Bayesian inference uses a set of hypotheses to infer the environment by changing the probability of the hypotheses, the hypotheses are required to be sufficiently different from each other that each is sensitive to the

environment. Therefore, each hypothesis $h_i$ is required to have a sharp peak in the diagonal element datum $d_i$. If the effective domain of the data is a subset of the set of all data, the likelihood of a hypothesis must be contained in the effective domain. This is simply an extension of Bayesian inference, and that is why it is called an excess Bayesian process. Therefore, we formulate the abovementioned excess Bayesian process as follows: Given a joint probability, $P^t\left(d_i, h_j\right), i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$, a binary relation is obtained by a given threshold value $\theta$:

$$\left(h_j, d_i\right) \in R^t, if P^t\left(d_i, h_j\right) > \theta; \left(h_j, d_i\right) \notin R^t, \text{otherwise} \quad (15)$$

The detection of data is defined by taking a diagonal element $\left(h_i, d_i\right)$ and considering inverting the relation either from $\left(h_i, d_i\right) \in R^t$ to $\left(h_{i+1}, d_{i+1}\right) \notin R^t$ or from $\left(h_i, d_i\right) \notin R^t$ to $\left(h_{i+1}, d_{i+1}\right) \in R^t$. This is expressed as a set $\mathbf{M}^t$ such that

$$\begin{aligned}\mathbf{M}^t = \{M_k = \left(h_i, d_i\right) \Big| \Big(\left(h_i, d_i\right) \in R^t \wedge \left(h_{i+1}, d_{i+1}\right) \notin R^t\Big) \\ \vee \Big(\left(h_i, d_i\right) \notin R^t \wedge \left(h_{i+1}, d_{i+1}\right) \in R^t\Big)\Big\}\end{aligned} \quad (16)$$

where if $M_k = \left(h_i, d_i\right)$ and $M_{k+1} = \left(h_j, d_j\right)$, then $i < j$. When the Bayesian inference is expressed as Equations 3, 4, some data are ignored, which decreases the probability of $P^t\left(d_i, h_j\right)$ when $d_i$ is not found under the given circumstances. A set $\mathbf{M}^t$ assigns a subset of the relation of diagonal elements. By constructing the set $\mathbf{M}^t$, we contract the universe of discourse depending on experience. In non-Bayesian probability theory, the following holds:

$$P\left(h, d\right) = P\left(h|d\right) \sum_k P\left(d, h_k\right) \quad (17)$$

the following equation also holds:

$$P\left(d\right) = \sum_k P\left(d, h_k\right) = \sum_j P\left(d_j\right) \quad (18)$$

The contraction of the universe depending on experience is implemented as follows: we assume that according to the belief that the universe consists only of experiences, the probability of the data that have been experienced is 1.0. That is,

$$P\left(d\right) = \sum_k P\left(d, h_k\right) = 1.0 \quad (19)$$

Substituting Equation 19 with Equation 17 and introducing an iterative process as before, with a given time step $t$, the newly contracted universe considering the data that have been experienced is denoted at each time step by $d'$, and one obtains

$$P^{t+1}\left(h, d\right) = P^t\left(h|d\right) = \frac{P^t\left(h, d\right)}{P^t\left(d'\right)} \quad (20)$$

The probability of the contracted data, $P^t\left(d'\right)$, at each time step is calculated within the contracted universe of discourse by Equation 16, so we have

$$P^t\left(d'\right) = \sum_{j=p}^{q} P^t\left(h_j, d\right) \quad (21)$$

where the domain of summation of $P^t\left(h_j, d\right)$ is given by

$$\begin{aligned}p = \pi M_k, q = \pi M_{k+1} with\, M_k, M_{k+1} \in \mathbf{M}^t,\ k = 1,\ 2, \\ \ldots,\ m-1\end{aligned} \quad (22)$$

which assigns either the relation $\{\left(h_i, d_i\right) \in R^t\}$ that has been experienced or the relation $\{\left(h_i, d_i\right) \notin R^t\}$ that has not been experienced, where

$$\pi M_k = \pi\left(h_i, d_i\right) = i \quad (23)$$

which assigns the index of the diagonal element. For $M_m \in \mathbf{M}^t$, $M_m = (h_N,\ d_N)$, and then $p = \pi M_{m-1}, q = N$. In the *excess Bayesian procedure*, first, $P^t\left(d\right)$, as a part of the universe, is regarded as a universe, which entails $P^t\left(d\right) = 1.0$, which cancels any effect of $P^t\left(d\right)$. Then, in turn, the data that have been experienced are denoted by $d'$, and then $P^t\left(d'\right)$ is calculated from the small area assigned by Equation 16. The following mnemonic expression might help us understand the basis of the excess Bayesian inference:

$$P\left(h, d\right) = \frac{P\left(h, d\right)}{P\left(d\right)} P\left(d\right) \quad (24)$$

While Equation 24 is indeed trivial, $P\left(d\right)$ in the numerator on the right-hand side must be regarded as 1.0, and $P\left(d\right)$ in the denominator on the right-hand side must be regarded as $P^t\left(d'\right)$. Thus, we obtain Equation 20.

By using Equation 22 or 23, the joint probability is obtained by summation with respect to the hypotheses, so now it reads as:

$$P^{t+1}\left(h, d\right) = P^t\left(h|d\right) = \frac{P^t\left(h, d\right)}{P^t\left(d'\right)} = \frac{P^t\left(h, d\right)}{\sum_{j=p}^{q} P^t\left(h_j, d\right)} \quad (25)$$

Symmetrically, the joint probability is obtained by summation with respect to the data, so now it reads as:

$$P^{t+1}\left(h, d\right) = P^t\left(d|h\right) = \frac{P^t\left(h, d\right)}{P^t\left(h'\right)} = \frac{P^t\left(h, d\right)}{\sum_{j=p}^{q} P^t\left(h, d_j\right)} \quad (26)$$

Using Equations 25, 26, not only the joint probability in the assigned area but also the joint probability outside the assigned area is normalized. Therefore, the effect of Bayesian inference contributes to the area outside the originally assigned area. The algorithmic representation of excess Bayesian inference is shown in Appendix C.

Figure 3 shows how Equations 25, 26 play a role in calculating the joint probability. This is a result of numerical

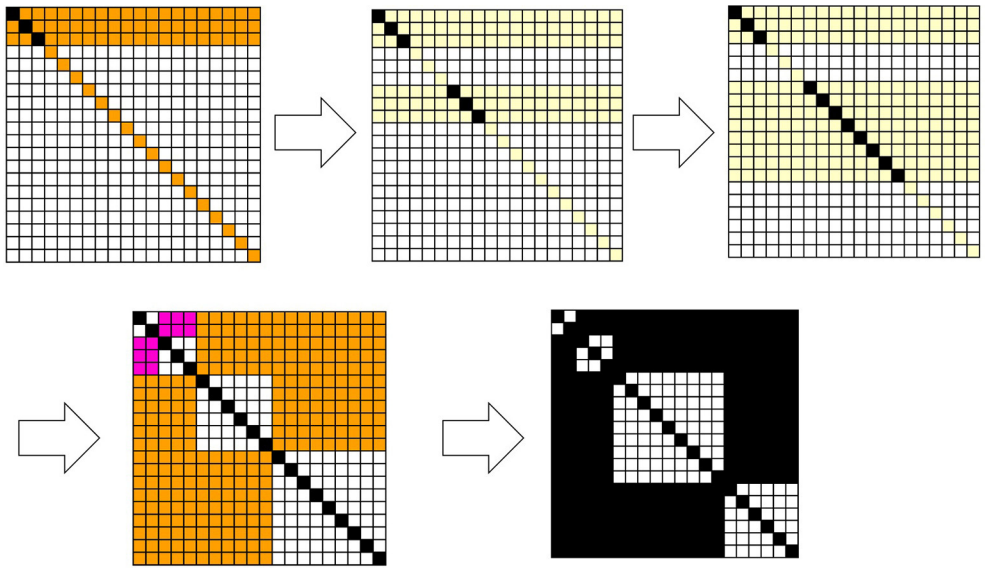


FIGURE 3
Snapshots of the joint probabilities of data and hypotheses and their corresponding relation $R$. The articulation of the data and hypotheses with respect to the diagonal elements is determined by Equation 22. The joint probabilities are colored from low to high in the order of white, pale yellow, orange, pink, and black. The last diagram in the sequence ordered by the arrows represents a binary relation. A black cell at $(h, d)$ represents $(h, d) \in R$, and a white cell represents $(h, d) \notin R$.

simulation in which after the classic Bayesian process is used in the first 30 steps, the newly introduced excess Bayesian process, defined by Equations 25, 26, is implemented. In Figure 3, $H = \{h_1, h_2, \ldots, h_{20}\}, D = \{d_1, d_2, \ldots, d_{20}\}$, and initially, $P^{t=0}(d_i|h_i) = 0.8$; in the case of $i \neq j$, $P^{t=0}(d_j|h_i) = \frac{1-0.8}{19} = 0.011$. The probability of each hypothesis is such that $P^{t=0}(h) = 0.05$. The time development follows Equations 3, 4, where a specific data value, $d$ in $P^t(h|d)$, is given at each time. In Figure 2, the specific $d$ is randomly given from a subset of $D$, such as $D - E$ with $E = (\{d_3, d_4, d_5, d_6\} \cup \{d_{15}, d_{16}, d_{17}, d_{18}, d_{19}, d_{20}\})$. Since the probability of $d$ is calculated cumulatively, as time proceeds, the probability converges to the situation; $P^t(d_u) = 0.0$ for $d_u \in E$ , and $P^t(d_s)$ =1/10=0.10 for $d_s \notin E$. However, Bayesian inference is adopted only for 30 time steps. If at time step $t$, $d_s \notin E$ is given, then $P^t(h|d_s) = \frac{P^t(d_s|h)P^t(h)}{\sum_k P^t(d_s|h_k)P^t(h_k)}$ is calculated for any $h$. From this, $P^{t+1}(h) = P^t(h|d_s)$ is obtained by Equation 9. For the first 30 time steps, $P^{t+1}(d, h)$ is calculated by $P^{t+1}(d)P^{t+1}(h)$. Each matrix in Figure 3 is obtained as a heatmap, in which if $P^{t+1}(d, h) \geq 0.01$, the cell is painted black; if $0.01 > P^{t+1}(d, h) \geq 0.008$, it is painted pink; if $0.008 > P^{t+1}(d, h) \geq 0.002$, it is painted orange; if $0.002 > P^{t+1}(d, h) \geq 0.0006$, it is painted pale yellow; and otherwise, it is painted white.

In the excess Bayesian process, $\mathbf{M}^t$ is determined to assign data that have been experienced. When an element of $\mathbf{M}^t = \{M_1 = (h_1, d_1), M_2 = (h_3, d_3), M_3 = (h_6, d_6), M_4 = (h_{12}, d_{12})\}$ is chosen, one obtains $p = \pi M_k, q = \pi M_{k+1}$. As shown in Figure 3, a square whose vertices are $(h_p, d_p), (h_p, d_{q-1}), (h_{q-1}, d_p), (h_{q-1}, d_{q-1})$ becomes the diagonal relation that is defined by Equation 15. Squares are added due to this diagonal relation, such as the square whose vertices are $\{(h_1, d_1), (h_1, d_2), (h_2, d_1), (h_2, d_2)\}$, $\{(h_3, d_3), (h_3, d_5), (h_5, d_3), (h_5, d_5)\}$, $\{(h_6, d_6), (h_6, d_{11}), (h_{11}, d_6), (h_{11}, d_{11})\}$, and $\{(h_{12}, d_{12}), (h_{12}, d_{20}), (h_{20}, d_{12}), (h_{20}, d_{20})\}$, and consists of $(h_i, d_j) \in R$. The above example demonstrates how the joint probabilities of data and hypotheses in the area assigned by the data that have been experienced and their corresponding hypotheses are normalized by dividing those probabilities by $\sum_{j=p}^{q} P^t(h_j, d)$ and $\sum_{j=p}^{q} P^t(h, d_j)$. Finally, each cell in the relation is painted black if $P^{t+1}(d, h) \geq 0.0006$; otherwise, it is painted white.

This normalization procedure enhances the diagonal relationship between the data and hypotheses beyond the experienced domain, which implies an explicit one-to-one correspondence between the data and hypotheses. In addition, the joint probabilities outside the diagonal relation are no longer negligibly small values. In Bayesian inference, only a diagonal relation is accessible as obtained, and any pair of data and hypotheses outside a diagonal relation are necessarily ignored; therefore, their joint probabilities almost disappear. In contrast, during the *excess Bayesian inference* process, pairs of data and hypotheses outside the diagonal relation are also enhanced, and the corresponding joint probabilities are increased. Although this might seem paradoxical and counterintuitive, it is indeed

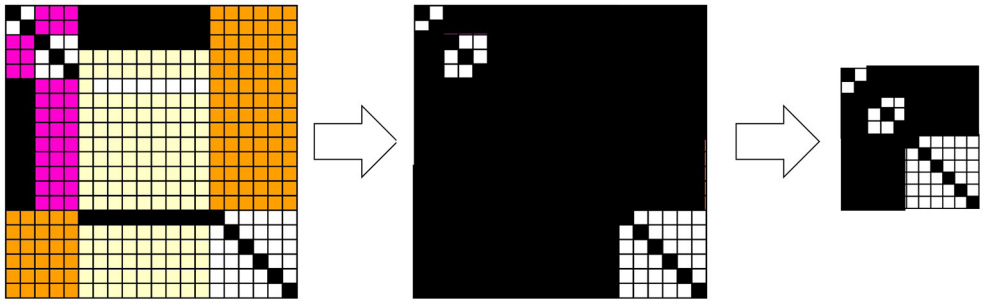


FIGURE 4
Snapshots of the joint probabilities of data and hypotheses and their corresponding relation $R$. The articulation of the data and hypotheses with respect to the diagonal elements is determined by Equation 23. The joint probabilities are colored from low to high in the order of white, pale yellow, orange, pink, and black (left). A black cell at $(h, d)$ represents $(h, d) \in R$, and a white cell represents $(h, d) \notin R$ (center right).

true that overestimating one's own experience allows alternative possibilities to emerge outside the given experience.

We previously showed that the psychological origin of quantum mechanics and/or quantum logic results not from a poor choice of basis vectors but from generating a context in which an object outside a brain is uniquely connected with a representation inside a brain, where outside the context, the object is connected to all images and the image is connected to all objects (Gunji and Haruna, 2022; Gunji and Nakamura, 2022a,b). In other words, there is a one-to-one relation within each context, and outside the context (i.e., background), objects are connected to all representations and vice versa. Since a one-to-one relation entails a Boolean algebra, this system entails multiple Boolean algebras connecting *via* the least and the greatest elements resulting from the background. This is the implementation of quantum logic without a Hilbert space. In this paper, we added another explanation for the psychological origin of quantum logic by introducing the excess Bayesian process.

The results from a numerical simulation of the excess Bayesian inference are shown in Figure 3, where the domain assigned by data that have been experienced is determined by Equation 22, which articulates the relationship between data that have been experienced and data that have not been experienced, and through this articulation, it assigns the vertex of the diagonal relation. Finally, the joint probabilities of the data and hypotheses entail, by induction, certain binary relations consisting of multiple diagonal relations and newly formed relations outside the diagonal relations, where pairs $(h_i, d_j)$ outside the diagonal relations remain in their corresponding relations $R$.

Figure 4 shows the results from numerical simulations of the excess Bayesian inference, where the domain assigned by the experienced data is determined by Equations 22, 23, in which only data that have been experienced are assigned as articulated. Therefore, diagonal elements whose data have been experienced constitute a diagonal relation, while those whose data have not been experienced constitute the background outside the diagonal relations. Similar to Figure 3, any $(h_i, d_j)$ outside the diagonal relations is in $R$. The central diagram in Figure 4 shows a large area consisting of cells representing $(h, d) \in R$, which contains diagonal cells. In this relation, for some $d$, any $h$ is in a relation such that $(h, d) \in R$, and for some $h$, any $d$ is in a relation such that $(h, d) \in R$. These pairs of $(h, d) \in R$ can be canceled out with respect to a given rough-set lattice approximation because for some $d$ such that any $h$ has a relation to it, $D_*\left(H^*\left(\{d, \ldots\}\right)\right) = D_*(H) = D$ is not an element of a lattice. Thus, a $d$ such that any $h$ has a relation to it and an $h$ such that any $d$ has a relation to it can be mutually removed from a relation $R$. The diagram on the right in Figure 4 shows such a relation, where the redundant rows and columns have been removed.

Figures 5, 6 show a corresponding lattice obtained from the relation in Figure 3, where this lattice is defined by Equations 7–9. It is easy to see that they are the same kind of orthomodular lattices that correspond to quantum logic. Figure 5 shows a relation between the data and the hypotheses and their corresponding sublattices. When we focus on the $3 \times 3$ diagonal relation between $\{d_{21}, d_{22}, d_{23}\}$ and $\{h_{21}, h_{22}, h_{23}\}$, the it is easy to see that

$$D_*\left(H^*\left(\{d_{21}, d_{22}\}\right)\right) = D_*\left(\{h_{21}, h_{22}\} \cup \overline{\{h_{21}, h_{22}, h_{23}\}}\right)$$
$$= D - \left(\overline{\{d_{21}, d_{22}\}} \cup \{d_{23}\}\right) = \{d_{21}, d_{22}\}$$

and that for a subset of $D$, $X$, consisting of elements belonging to different diagonal relations, Equation 12 does not hold, since $D_*\left(H^*(X)\right) = D$.

Let us also note here that $D_*\left(H^*\left(\{d_{21}, d_{22}, d_{23}\}\right)\right) = D$. Therefore, this $3 \times 3$ diagonal relation yields the power set of $\{d_{21}, d_{22}, d_{23}\}$ except for $\{d_{21}, d_{22}, d_{23}\}$. Thus, if the greatest element is represented by $D$, the $3 \times 3$ diagonal relation entails a $2^3$-Boolean lattice. These considerations are general; therefore, an $n \times n$ diagonal relation entails a $2^n$-Boolean lattice.

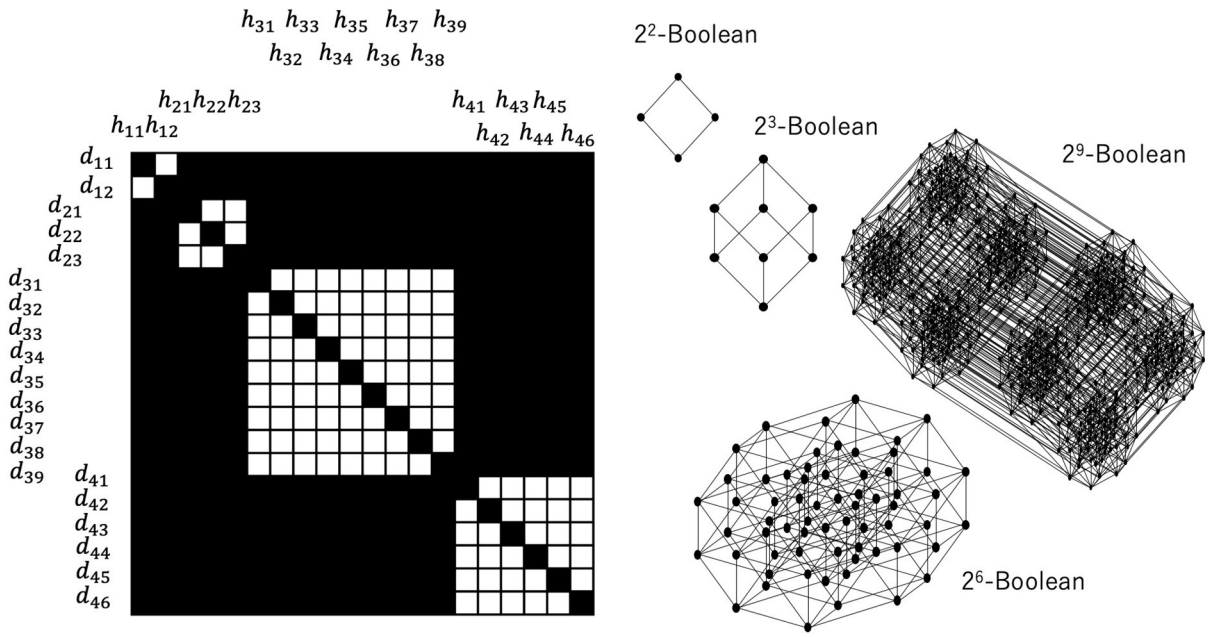



FIGURE 5
Relation between data and hypotheses consisting of $(2 \times 2)$, $(3 \times 3)$, $(9 \times 9)$, $(6 \times 6)$ diagonal relations and their corresponding Boolean lattices. All relations outside the diagonal relations constitute the greatest element, which fuses all the other Boolean lattices.

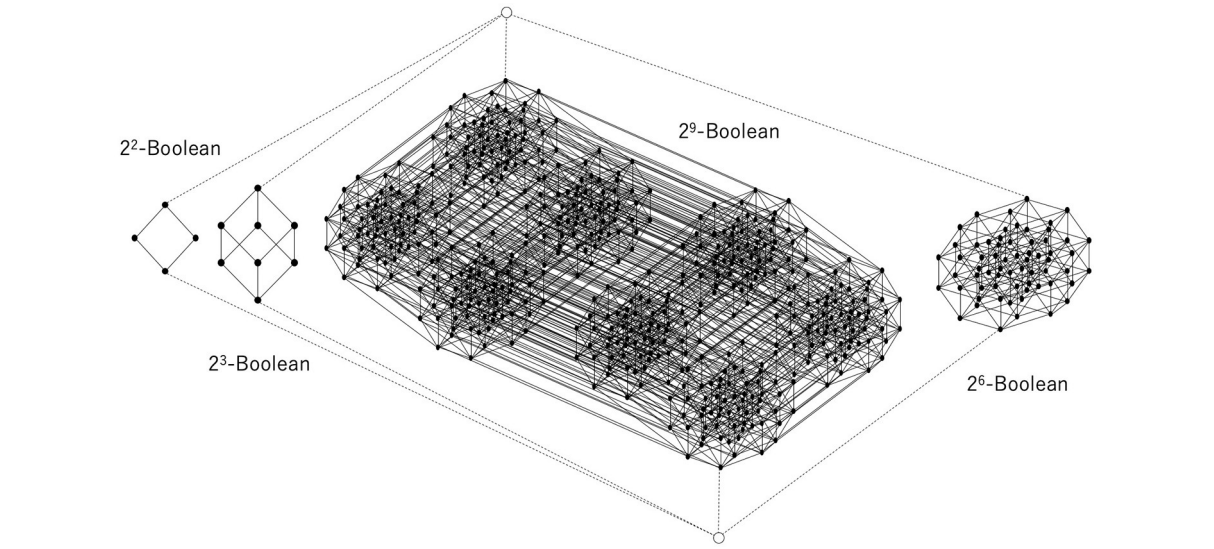



FIGURE 6
Disjoint union of some Boolean lattices, the least and greatest elements of which are common to all Boolean lattices. This is obtained from the relation shown in Figure 5.

Figure 6 shows the whole construction of a lattice corresponding to a whole relation. Each black circle represents an element of a (sub)lattice that is a subset of $D$. The greatest element and the least element are represented by white circles, and they are connected to the least and the greatest element of each Boolean lattice by broken lines, which implies that the greatest element of each Boolean lattice is the same as $D$ and that the least element of each Boolean lattice is the empty set.

Thus, one can say that the whole structure is a disjoint union of multiple Boolean (sub)lattices except for the least and the greatest element. It can be straightforwardly verified that this is an orthomodular lattice of the kind that is well known from quantum logic (see Appendix A).

Therefore, classical Bayesian inference gives rise to a Boolean lattice, while *excess Bayesian inference* gives rise to an orthomodular lattice. This is because classical Bayesian inference ignores the outside of the diagonal relation, which leads to a Boolean lattice. This implies that a decision-maker who applies a classical logical reductionistic approach to restricted pairs of data and hypotheses that are being experienced ignores the pairs of data and hypotheses that are not being experienced. In contrast, a decision-maker who not only uses classical Bayesian inference but augments it *via* an *excess Bayesian inference* process can enhance the basic one-to-one correspondence between the data and hypotheses by normalization within the domain of experience. This leads to multiple sub-relations, which are diagonal relations and relations that constitute the background of the diagonal relations. If the background consists of no relation [i.e., $(h, d) \notin R$], multiple (sub)diagonal relations constitute the one diagonal relation expressing the Boolean algebra. However, if the background consists of relations [i.e., $(h, d) \in R$], this same background plays a key role in constituting the common greatest element by which multiple Boolean lattices are fused.

Since any element except for the least and greatest elements has a unique complement in each Boolean sublattice, any element can be regarded as an orthocomplement. Indeed, if an element and its orthocomplement in the lattice, such as those shown in Figure 6, can be compared with each other and with respect to the order, they are both in the same Boolean lattice participating as a sublattice, and then the distributive law holds for the element and its orthocomplement. This is why the orthomodular law holds for that lattice. Therefore, the result is that excess Bayesian inference can entail and accommodate quantum logic.

## Discussion and conclusion

We initially investigated the claim that the cognitive perspective based on the free energy principle could seemingly conflict with quantum cognition since the former tends toward optimization by removing redundant search space and the latter tends toward ambiguous and non-optimal decision-making. Since the free energy principle mathematically and formally includes classical Bayesian inference as an instance, one can estimate *via* optimization techniques the distribution of the joint probability of data and hypotheses, and this can be expressed as a binary relation. If the data and hypotheses are replaced by objects outside the brain and representations (or "images") inside the brain, respectively, and if symmetry between objects and images is assumed, one can logically evaluate the hidden structure between objects and representations, "images," with respect to a Boolean lattice structure. In this sense, one can estimate how a Boolean lattice (i.e., classical logic) and an orthomodular lattice (i.e., quantum logic) can arise from a given inference system. We also examined how these considerations can bridge the considerations of the free energy principle with those of quantum cognition. After a reviewer's comment (we thank them for this remark), it has come to our attention that bipolar fuzzy relations (Zhang, 1998, 2021a,b) should be considered and that comparing our previously proposed rough set approximation with respect to quantum cognition might enhance its scope by enabling a way to connect Bayesian inference and causal inference. Although this new development is beyond the scope of the present study, we maintain an interest in future investigations that could shed some light on the epistemological and ontological bases of quantum cognition and cognition at large.

Applying the free energy principle consists of minimizing the difference between the conditional probability under a given experience and the marginal probability *via* the minimization of the prediction error. Compared to cortical processes, in predictive coding, the former process is considered a classical Bayesian inference that has to be interpreted as a hierarchical top-down process, and the latter process is interpreted as a hierarchical bottom-up process. This is based on the premise that when new data are received and prediction errors are detected, the prediction errors lead to a modification of the prediction system. In other words, top-down Bayesian inference can make a system "see" new data through the old filters based on previous experiences. Thus, a system is subject to unavoidably becoming stuck in previous experience. Moreover, active inference (embodied cognition) is also a top-down process that enhances the Bayesian inference process. To receive data that are consistent with previous data, the top-down process makes the body move and act accordingly. This leads to stubborn inference based on experience. In this sense, an active inference might deploy a certain kind of possible excess Bayesian inference process. In other words, the active inference is a flexible interface based on a body between the environment and a stubborn inference system. This has been described as embodied intelligence.

In our work, we propose that excess Bayesian inference plays a key role in the process of cognition, much greater than that of active inference. This is because classical Bayesian inference restricts the domain of joint probabilities of data and hypotheses, while excess Bayesian inference realizes and enforces a one-to-one correspondence between data and hypotheses beyond the initially restricted domain. Therefore, this is called *excess* Bayesian inference. Our proposed process is implemented *via*

a stepwise, iterative renormalization of the joint probability at each step divided by the sum of all joint probabilities in a restricted domain. This renormalization is achieved with respect to data and hypotheses independently, and these processes are not performed simultaneously. For any data, the joint probability is renormalized with respect to the hypotheses, and for any hypotheses, the joint probability is renormalized with respect to the data. Therefore, the effect of renormalization, which is inherent in the process of excess Bayesian inference, can influence not only the initially restricted domain but also the domains outside the initial domain. This implies that excess Bayesian inference can bring out non-zero joint probabilities that are noted as significant even outside the initially restricted domain, a potential that classical Bayesian inference lacks.

The distribution of the joint probabilities of data and hypotheses is expressed as a binary relation if a threshold value is introduced to distinguish a relation from the lack of a relation. Although Bayesian inference is expressed as a simple diagonal relation, excess Bayesian inference is expressed as multiple (sub)diagonal relations whose backgrounds consist of relations. The relations between the data and hypotheses are transformed into an algebraic structure called a lattice, and one can estimate the differences between classical Bayesian inference and the newly proposed *excess Bayesian inference* in terms of the lattice structure, each reflecting its underlying logical structure. Classical Bayesian inference is expressed as a simple Boolean lattice (classical logic), and *excess Bayesian inference* is expressed as an orthomodular lattice (quantum logic). This, in turn, implies that excess Bayesian inference can bridge the Boolean lattice structure of the classic Bayesian inference by encompassing a wider orthomodular lattice similar to those of quantum logic. From these considerations, we conclude that the basis of quantum cognition results from a radical extension of the Bayesian inference framework rather than simply alternative versions of classic Bayesian inference and that the free energy minimization principle can be bridged with quantum cognition *via* the proposed excess Bayesian inference scheme.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

Y-PG and SS constructed the idea of excess Bayesian inference and VB contributed on the development of ideas. Y-PG and VB wrote the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by JSPS Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences, JPJS00120351748. VB acknowledges partial support from the Italian CNR Foresight Institute's STEM materials program and from the Service de Physique des Systèmes Complexes et Mécanique Statistique of ULB.

## Acknowledgments

VB wishes to thank Dr. Pier-Francisco Moretti of CNR for his encouragement and exchange of ideas on information, meaning, and action.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2022.910161/full#supplementary-material

## References

Aerts, D. (2009). Quantum structure in cognition. *J. Math. Psychol.* 53, 314–348. doi: 10.1016/j.jmp.2009.04.005

Aerts, D., Arguëlles, J. A., Beltran, L., Geriente, S., and Sassoli, D. E., Bianchi, M., Sozzo, S., Veloz, T. (2019). Quantum entanglement in physical and cognitive systems: a conceptual analysis and a general representation. *Eur. Phys. J. Plus* 134, 493. doi: 10.1140/epjp/i2019-12987-0

Aerts, D., Gabora, L., and Sozzo, S. (2013). Concepts and their dynamics: a quantum-theoretic modeling of

human thought. *Top. Cogn. Sci.* 5, 737–772. doi: 10.1111/tops.12042

Aerts, D., Broekaert, J., Gabora, L., and Veroz, T. (2012). The guppy effect as interference. *Quant. Int.*, 7620, 36–74. doi: 10.1007/978-3-642-35659-9_4

Allen, M., and Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese* 195, 2459–2482. doi: 10.1007/s11229-016-1288-5

Arecchi, F. T. (2011). Phenomenology of consciousness from apprehension to judgment. *Nonlinear Dyn. Psychol. Life Sci.* 15, 359–375.

Arecchi, F. T. (2003). Chaotic neuron dynamics, synchronization and feature binding: quantum aspects. *Mind Matter* 1, 15–43.

Asano, M., Khrennikov, A., Ohya, M., Tanaka, Y., and Yamato, I. (2015). *Quantum Adaptivity in Biology: From Genetics to Cognition*. Berlin: Springer.

Basios, V., and Gunji, Y. P. (2021). "Chaos, rhythms and processes in structure and function: extending Bayesian Inference," in *Proceedings of the Science and Technology Foresight Workshop, "A Quest for An Interface Between Information and Action"*, Edited by Pier Francesco Moretti and Vasileios Basios. Available online at: http://www.foresight.cnr.it/pubblications/issn.html (accessed August 23, 2022).

Bruza, P. D., Wang, Z., and Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. *Trends Cogn. Sci.* 19, 383–393. doi: 10.1016/j.tics.2015.05.001

Busemeyer, J. R., and Bruza, P. D. (2012). *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press.

Çatal, O., Verbelen, T., Maele, D. E., Dhoedt, T. V., and Safron, B. A. (2021). Robott navigation as hierarchical active inference. *Neural Networks* 142, 192–204. doi: 10.1016/j.neunet.2021.05.010

Dzhafarov, E. N., Kujala, J. V., and Cervantes, V. H. (2016). "Contextuality-by-default: a brief overview of ideas, concepts, and terminology," in *Lecture Notes in Computer Science*, eds H. Atmanspacher, T. Filk, E. Pothos (Cham: Springer), 12–23.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., and Kiebel, S. (2009a). Cortical circuits for perceptual inference. *Neural Networks* 22, 1093–1104. doi: 10.1016/j.neunet.2009.07.023

Friston, K. J., and Kiebel, S. (2009b). Predictive coding under the free-energy principle. *Philos. Trans. Royal Soc. B. Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K. J., and Kilner, J., and Harison, L. (2006). A free energy principle for the brain. *J. Physiol.* 1000, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Gunji, Y.-., P., Kawai, T., Murakami, H., Tomaru, T., Minoura, M., et al. (2021). Lévy walk in swarm models based on Bayesian and inverse bayesian inference. *Comput. Struct. Biotechnol. J.* 19, 247–260. doi: 10.1016/j.csbj.2020.11.045

Gunji, Y.-P., and Haruna, T. (2010). A non-boolean lattice derived by double indiscernibility. *Trans. Rough Sets* 12, 211–225. doi: 10.1007/978-3-642-14467-7_11

Gunji, Y. P., Murakami, H., and Tomaru, T., and Vasios, V. (2018). Inverse bayesian inference in swarming behavior of soldier crabs. *Philos. Trans. Royal Soc. A.* 376, 20170370. doi: 10.1098/rsta.2017.0370

Gunji, Y. P., and Nakamura, K. (2022a). Psychological origin of quantum logic:an orthomodular lattice derived from natural-born intelligence without hilbert space. *BioSystems* 215, 104649. doi: 10.1016/j.biosystems.2022.104649

Gunji, Y. P., and Nakamura, K. (2022b). "Kakiwari: The device summoning creativity in art and cognition," in *Unconventional Computing, Philosophies and Art*, ed A. Adamatzky (Singapore: World Scientific), 135–168.

Gunji, Y. P., Nakamura, K., Minoura, M., and Adamatzky, A. (2020). Three types of logical structure resulting from the trilemma of free will, determinism and locality. *BioSystems* 195, 104151. doi: 10.1016/j.biosystems.2020.104151

Gunji, Y. P., and Shinohara, S., Haruna, T., Basios, V. (2017). Inverse bayesian inference as a key of consciousness featuring a macroscopic quantum logic structure. *BioSystems* 152, 44–63. doi: 10.1016/j.biosystems.2016.12.003

Gunji, Y. P., and Sonoda,K., Basios, V. (2016). Quantum cognition based on an ambiguous representation derived from a rough set approximation. *Biosystems* 141, 55–66. doi: 10.1016/j.biosystems.2015.12.003

Gunji, Y. P., and Haruna, T. (2022). Concept formation and quantum-like probability from nonlocality in Cognition. *Cognit. Psychol.* 14, 1328–1349. doi: 10.1007/s12559-022-09995-1

Haven, E., and Khrennikov, A. (2013). *Quantum Social Science*. Cambridge: Cambridge University Press.

Heunen, C., and Vicary, J. (2019). *Categories for Quantum Theory, An introduction*. Oxford: Oxford University Press.

Ishwarya, M. S., and Kumar, C. H. A. (2020a). Quantum aspects of high dimensional formal representation of conceptual spaces: a model for achieving consciousness. *Cognit. Comput.* 12, 563–576. doi: 10.1007/s12559-020-09712-w

Ishwarya, M. S., and Kumar, C. H. A. (2020b). Decision-making in cognitive paradoxes with contextuality and quantum formalism, applied soft. *Computing* 95, 106521. doi: 10.1016/j.asoc.2020.106521

Khrennikov, A. (2001). Linear representations of probabilistic transformations induced by context transitions. *J. Phys. A.* 34, 9965–9981. doi: 10.1088/0305-4470/34/47/304

Khrennikov, A. (2010). *Ubiquitous Quantum Structure: From Psychology to Finances*. Berlin: Springer.

Khrennikov, A. (2021). Quantum-like model for unconscious-conscious interaction and emotional coloring of perceptions and other conscious experiences. *Biosystems* 208, 104471. doi: 10.1016/j.biosystems.2021.104471

Kumar, C. H. A., Ishwarya, M. S., and Loo, C. K. (2015). Formal concept analysis approach to cognitive functionalities of bidirectional associative memory. *Biol. Insp. Cognit. Architec.* 12, 20–33. doi: 10.1016/j.bica.2015.04.003

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Front. Robot. AI.* 5, 21. doi: 10.3389/frobt.2018.00021

Ozawa, M., and Khrennikov, A. (2021). Modeling combination of question order effect, response replicability effect, and QQ-equality with quantum instruments. *J. Math. Psych.* 100, 102491. doi: 10.1016/j.jmp.2020.102491

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. B. Biol. Sci.* 371, 19. doi: 10.1098/rstb.2016.0007

Seth, A. K., and Tsakiris, M. (2018). Being a beast machine: the somatic basis of selfhood. *Trends Cognit. Sci.* 22, 969–981. doi: 10.1016/j.tics.2018.08.008

Shivahare, R., and Cherukuri, A. K. (2017). Three-way conceptual approach for cognitive memory functionalities. *Int. J. Mach. Learn. Cybernetic.* 8, 21–34. doi: 10.1007/s13042-016-0593-0

Varela, F. (1997). Patterns of life: intertwining identity and cognition. *Brain Cognit.* 34, 72–87. doi: 10.1006/brcg.1997.0907

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind*. Cambridge, MA: MIT Press.

Walsh, K. S., McGovern, D. P., Clark, A., and O'Cornnell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annal. N.Y. Acad. Sci.* 1464, 242–268. doi: 10.1111/nyas.14321

Yao, Y. Y. (2004). "A comparative study of formal concept analysis and rough set theory in data analysis," in *RSCTC 2004. LNCS (LNAI), 3066,* eds S. Tsumoto, R. Slowinski, J. Komorowski, and J. W. Grzymala-Busse (Heidelberg: Springer), 59–68.

Yon, D., Lange, D. E., and Press, F. P. C. (2019). The predictive brain as a stubborn scientist. *Trends Cognit. Sci.* 23, 6–8. doi: 10.1016/j.tics.2018.10.003

Zhang, W. R. (2021a). Ground-0 Axioms vs. first principles and second law: From the geometry of light and logic of photon to mind-light-matter unity-AI&QI. *IEEE/CAA J. Automatic. Sinica.* 8, 534–553. doi: 10.1109/JAS.2021.1003868

Zhang, W. R. (2021b). Science vs. sophistry—a historical debate on bipolar fuzzy sets and equilibrium-based mathematics for AI&QI. *J. Int. Fuzzy Syst.* 6781–6799. doi: 10.3233/JIFS-210692

Zhang, W. (1998). "(Yin) (Yang) bipolar fuzzy sets," in *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228)* (Anchorage, AK), 1, 835–840.

frontiers | Frontiers in Neurorobotics

TYPE Hypothesis and Theory
PUBLISHED 30 September 2022
DOI 10.3389/fnbot.2022.846982

Check for updates

OPEN ACCESS

EDITED BY
Inês Hipólito,
Humboldt University of Berlin,
Germany

REVIEWED BY
Tetsushi Nonaka,
Kobe University, Japan
Riccardo Manzotti,
Università IULM, Italy

*CORRESPONDENCE
Vadim Weinstein
vadim.weinstein@oulu.fi
Steven M. LaValle
steven.lavalle@oulu.fi

RECEIVED 31 December 2021
ACCEPTED 28 July 2022
PUBLISHED 30 September 2022

CITATION
Weinstein V, Sakcak B and LaValle SM (2022) An enactivist-inspired mathematical model of cognition. *Front. Neurorobot.* 16:846982. doi: 10.3389/fnbot.2022.846982

COPYRIGHT
© 2022 Weinstein, Sakcak and LaValle. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# An enactivist-inspired mathematical model of cognition

Vadim Weinstein*, Basak Sakcak and Steven M. LaValle*

Center for Ubiquitous Computing, Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland

In this paper we start from the philosophical position in cognitive science known as enactivism. We formulate five basic enactivist tenets that we have carefully identified in the relevant literature as the main underlying principles of that philosophy. We then develop a mathematical framework to talk about cognitive systems (both artificial and natural) which complies with these enactivist tenets. In particular we pay attention that our mathematical modeling does not attribute contentful symbolic representations to the agents, and that the agent's nervous system or brain, body and environment are modeled in a way that makes them an inseparable part of a greater totality. The long-term purpose for which this article sets the stage is to create a mathematical foundation for cognition which is in line with enactivism. We see two main benefits of doing so: (1) It enables enactivist ideas to be more accessible for computer scientists, AI researchers, roboticists, cognitive scientists, and psychologists, and (2) it gives the philosophers a mathematical tool which can be used to clarify their notions and help with their debates. Our main notion is that of a sensorimotor system which is a special case of a well studied notion of a transition system. We also consider related notions such as labeled transition systems and deterministic automata. We analyze a notion called *sufficiency* and show that it is a very good candidate for a foundational notion in the "mathematics of cognition from an enactivist perspective." We demonstrate its importance by proving a uniqueness theorem about the minimal sufficient refinements (which correspond in some sense to an optimal attunement of an organism to its environment) and by showing that sufficiency corresponds to known notions such as sufficient history information spaces. In the end, we tie it all back to the enactivist tenets.

KEYWORDS

**enactivism, transition systems, automaton, cognitive modeling, information spaces, robotics**

## 1. Introduction: Mathematizing enactivism

The premise of this paper is to lay down a logical framework for analyzing agency in a novel way, inspired by enactivism. Classically, mathematical and logical models of cognition are in line with the cognitivist paradigm in that they rely on the notion of symbolic representation and do not emphasize embodiment or enactment (Newell and Simon, 1972; Fodor, 2008; Gallistel and King, 2009; Rescorla, 2016).

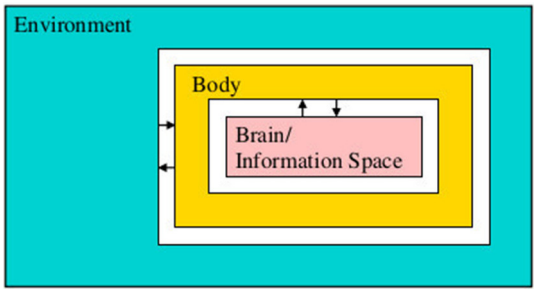



FIGURE 1
The environment, body, and nervous system (or brain) will be modeled as inseparable parts of a coupled transition system.

Cognitivism presumes that the world possesses objective structure and the contentful information of this structure is acquired and represented by the cognitive agent. This aligns well with the classical model-theoretic paradigm. In this paradigm a formal language is describing a static model (such as when sentences in the language of rings describe algebraic structures—such as rings).

In the cognitivist analogy, the agent possesses ("in its head") formulas of the language and the model is the world or the environment of the agent. If the formulas possessed by the agent hold in the model, then the agent's representation of the world is correct; otherwise, it is incorrect. Such view of cognitive agency is rejected by the enactivists either weakly or strongly depending on the branch of enactivism. For example, radical enactivism (Hutto and Myin, 2012, 2017) rejects this view strongly. Our question for this paper is: What would the mathematical logic of cognition look like, if even the radical enactivists were to accept it?

We do not take part in the cognitivist-enactivist, or the representationalist-antirepresentationalist debate (Pezzulo et al., 2011; O'Regan and Block, 2012; Gallagher, 2018; Fuchs, 2020). Rather, we take a somewhat extreme enactivist and antirepresentational view as our axiomatic starting point and as a theoretical explanatory target. Then we develop a mathematical theory that attempts to account for cognition in a way congruent with this view. Even though most forms of enactivism (even radical ones) have room for representation, it is not our main goal at the moment to bridge the gap between "basic minds" and "scaffolded minds," to use terminology of (Hutto and Myin, 2017). Thus, in this terminology, we are going to explore a mathematical (only) of *basic minds*.

The following "axioms" we take as fundamentals for our work:

(EA1) Embodiment. "From a third-person perspective the organism-environment is taken as the explanatory unit" (Gallagher, 2017). The environment, the body, and the nervous system are inseparable parts of the system which they form by coupling; see Figure 1. They cannot be meaningfully understood in isolation from each other. "Mentality is in all cases concretely constituted by, and thus literally consists of, the extensive ways in which organisms interact with their environments, where the relevant ways of interacting involve, but are not exclusively restricted to, what goes on in brains" (Embodiment Thesis Hutto and Myin, 2012).

(EA2) Groundedness. The brain does not "acquire" or "possess" contentful states, representations, or manipulate semantic information in any other way. "Mentality-constituting interactions are grounded in, shaped by, and explained by nothing more, or other, than the history of an organism's previous interactions. Nothing other than its history of active engaging structures or explains an organism's current interactive tendencies." [Developmental-Explanatory Thesis (Hutto and Myin, 2012)].

(EA3) Emergence. The crucial properties of the brain-body-environment system from the point of view of cognition emerge from the embodiment, the brain-body-environment coupling, the situatedness, and the skills of the agent. The agent's and the environment's prior structure come together to facilitate new structure which emerges through the sensorimotor engagement. "[T]he mind and world arise together in enaction, [but] their manner of arising is not arbitrary" (i.e. it is structured) (Varela et al., 1992).

(EA4) Attunement. Agents differ in their ways of attunement and adaptation to their environments, and in the skills they have. A *skill* is a potential possibility to engage *reliably* in complex sensorimotor interactions with the environment (Gallagher, 2017).

(EA5) Perception. Sensing and perceiving are not the same thing. Perception arises from skillful sensorimotor activity. To perceive is to become better attuned to the environment. O'Regan and Noë (2001) and Noë (2004) "Perception and action, sensorium and motorium, are linked together as successively emergent and mutually selecting patterns." Varela et al. (1992).

The mathematics we use to capture those ideas is a mixture of known and new concepts from theoretical robotics, (non-)deterministic automata and transition systems theory, and dynamical systems (Goranko and Otto, 2007). It will also build upon the *information spaces* framework, introduced in LaValle (2006) as a unified way to model sensing, actuation, and planning in robotics; the framework itself builds upon earlier ideas such as dynamic games with imperfect information (von Neumann and Morgenstern, 1944; Başar and Olsder, 1995), control with imperfect state information (Kumar and Varaiya, 1986; Bertsekas, 2001), knowledge states (Lozano-Pérez et al., 1984; Erdmann, 1993), perceptual equivalence classes (Donald

and Jennings, 1991; Donald, 1995), maze and graph-exploring automata (Shannon, 1952; Blum and Kozen, 1978; Fraigniaud et al., 2005), and belief spaces (Kaelbling et al., 1998; Roy and Gordon, 2003).

Although information spaces refer to "information," they are not directly related to Shannon's *information theory* (Shannon, 1948), which came later than von Neumann's use of information in the context of sequential game theory. Neither does "information" here refer to content-bearing information. One important intuition behind the information in information spaces is that more information corresponds to narrowing down the space of possibilities (for example of future sensorimotor interactions).

The main mathematical concept of this paper is a *sensorimotor system* (SM-system), which is a special case of a transition system. Sensorimotor systems can describe the body-brain system, the body-environment system as well as other parts of the brain-body-environment system. Given two SM-systems they can be *coupled* to produce another (third) SM-system. Mathematically, the coupling operation is akin to a direct product. We introduce several notions that describe the coupling of the agent and the environment from an outside perspective (not from the perspective of the agent or the environment). The main notion is that of *sufficiency*. In some sense it guarantees that the coupling is of "high fidelity." It does not compare "internal" models of the agent to "external" states of affairs. Rather it asks whether the way in which the agent engages in sensorimotor patterns is well structured. The notion of *sufficiency* compares the sensorimiotor capacity of the agent *to itself* by asking whether the past sensorimotor patterns (in a given environment) determine reliably the future sensorimotor patterns. We then introduce several related notions. The *degree of insufficiency* is a measure by which various agents can be compared in their coupling versatility (Def 4.11). *Minimal sufficient refinement* is a concept that can be used in the most vivid ways to illustrate how the sensorimotor interaction "enacts" properties of the brain-body-environment system. The notion of minimal sufficient refinement ties together mathematics of sensorimotor systems and the philosophical ideas of emergence, structural coupling and enactment of the "world we inhabit" (cf. Varela et al., 1992); see Example 4.25. We prove the uniqueness of minimal sufficient refinements (Theorem 4.19) and point out their connection to the notions of bisimulation and sufficient information mappings. *Strategic sufficiency* is a mathematically more challenging concept, but has appealing properties in the philosophical and practical sense. A sensor mapping is strategically sufficient *for* some subset of the state space $G$, if that sensor can (in principle) be used by the agent to reach $G$ [1]. Again, any sensor mapping has minimal strategic refinements, but this time they are not unique. Different minimal refinements in this case can be thought of as different adaptations to the same environmental demands.

Mathematically, sufficiency is a relative concept to some known notions in theoretical computer science and robotics: that of *bisimulation* in automata and Kripke model theory (Goranko and Otto, 2007), and *sufficient information mappings* in information spaces theory (LaValle, 2006).

Minimal sufficient refinements lead to unique classifications of agent-environment states that "emerge" from the way in which the agent is coupled to the environment, not merely from the way the environment is structured on its own. Thus, the world is simultaneously objectively existing (from the global "god" perspective), but also "brought about" by the agent.

This should be enough to answer the two questions that, according to Paolo (2018), any embodied theory of cognition should be able to provide precise answers to: What is its conception of bodies? What central role do bodies play in this theory different from the roles they play in traditional computationalism?

Section 2 introduces the basics of transition and SM-systems, their coupling, and other mathematical constructs such as quotients. Section 3 illustrates the introduced notions with detailed examples. Section 4 introduces the notion of sufficiency, sufficient refinements, and minimal sufficient refinements. We will prove the uniqueness theorem for the latter and illustrate the notions in a computational setting. We will explore the importance of sufficiency and related notions for the enactivist way of looking at cognitive organization. Finally, Section 5 ties the mathematics back to the philosophical premises.

## 2. Transition systems and sensorimotor systems

At the most abstract level, the central concept for our mathematical theory is that of a *transition system*. This is a standard definition from automata theory (for instance Goranko and Otto, 2007):

**Definition 2.1.** A *transition system* is a triple $(X, U, T)$ where $X$ is the *state space* (mathematically it is just a set), $U$ is the set of names for outgoing transitions (another set), and $T \subseteq X \times U \times X$ is a ternary relation.

The intuitive interpretation of $(X, U, T)$ is that it is possible to transition from the state $x_1 \in X$ to the state $x_2 \in X$ via $u \in U$ iff $(x_1, u, x_2) \in T$. We use the notation $x_1 \xrightarrow{u} x_2$ to mean that $(x_1, u, x_2) \in T$. Our notion of transition system is often called a *labeled* transition system in the literature, because each potential transition has a name or label, $u \in U$. However, we drop the term "labeled" because in Section 2.5 we will introduce a version of transition systems in which the states are relabeled, thereby introducing a new kind of labeling. Note that when working with such transition

1 This idea of a reachable set G is the simplest way to formalize affordances.

systems as modeling agency, we are safely within the realm of the Developmental-Explanatory Thesis (EA2). The following definitions are standard (although we do not restrict $X$ to be finite):

**Definition 2.2.** Let $\mathcal{X} = (X, U, T)$ and $\mathcal{X}' = (X', U', T')$ be transition systems. An *isomorphism* is a bijective function $f: X \to X'$ such that for all $x_1, x_2 \in X$ and $u \in U$ we have $(x_1, u, x_2) \in T \iff (f(x_1), u, f(x_2)) \in T'$. A *simulation* is a relation $R \subseteq X \times X'$ such that for all $(x_1, x_1') \in R$, all $u \in U$ and all $x_2 \in X$, we have that if $(x_1, u, x_2) \in T$, then there exists $x_2' \in X'$ with $(x_1', u, x_2') \in T'$ and $(x_2, x_2') \in R$. A *bisimulation* is a relation $R$ such that both $R$ and $R^T = \{(y, x) : (x, y) \in R\}$ are simulations.

The notation $\mathcal{X} \cong \mathcal{X}'$ means that $\mathcal{X}$, $\mathcal{X}'$ are isomorphic, and $\mathcal{X} \sim \mathcal{X}'$ means that there is a bisimulation $R$ such that $X = \mathrm{dom}(R)$ and $X' = \mathrm{ran}(R)$. We speak of *automorphism* and *autobisimulation*, if $\mathcal{X} = \mathcal{X}'$.

We are ready to make the first observation:

**Proposition 2.3.** *If $\mathcal{X} \cong \mathcal{X}'$, then $\mathcal{X} \sim \mathcal{X}'$.*

*Proof:* Let $f$ be an isomorphism $f: X \to X'$. Then $R = \{(x_1, x_2) \in X \times X' \mid x_2 = f(x_1)\}$ is a bisimulation. □

## 2.1. Transition systems as a unifying concept

There are several ways in which transition systems and their relatives appear in the literature relevant to us.

**Examples 2.4.** Let $(X, U, T)$ be a transition system.

1. Let $x_0 \in X$ and $F \subseteq X$. Let $\hat{T}: X \times U \to \mathcal{P}(X)$ be defined by $\hat{T}(x, u) = \{x_2 \in X \mid x_1 \xrightarrow{u} x_2\}$. Then $(X, U, \hat{T}, x_0, F)$ is a *nondeterministic automaton*. If in addition $X$ and $U$ are finite, then it is a *nondeterministic finite automaton* (NFA).
2. Let $\tilde{T}: X \times X \to \mathcal{P}(U)$ be the function $\tilde{T}(x_1, x_2) = \{u \in U \mid x_1 \xrightarrow{u} x_2\}$. Then $\tilde{T}(x_1, x_2)$ is the set of all $u$ that take $x_1$ to $x_2$. Then, $(X, \tilde{T})$ is a labeled directed graph in which the labels are subsets of $U$. Another way to think of it is as a labeled directed multigraph: the multiplicity of the edge from $x_1$ to $x_2$ is $n = |\tilde{T}(x_1, x_2)|$ and these $n$ edges are labeled by the labels from the set $\tilde{T}(x_1, x_2)$.
3. If for all $x_1 \in X$ and $u \in U$ there is a unique $x_2 \in X$ with $x_1 \xrightarrow{u} x_2$, let $\tau: X \times U \to X$ be the function defined such that $\tau(x_1, u) = x_2$ iff $x_1 \xrightarrow{u} x_2$. Let $x_0 \in X$ and $F \subseteq X$. Then $(X, U, \tau, x_0, F)$ is a *deterministic automaton*, and if $X$ and $U$ are finite, then it is a *deterministic finite automaton* (DFA). Without $F$, $(X, U, \tau, x_0)$ also satisfies the definition of the *temporal filter* of LaValle (2012, 4.2.3). In this case $X$ is the *information space* or the *I-space* (usually denoted by $\mathcal{I}$ instead of $X$), and $U$ is the observation space (usually denoted by $Y$ instead of $U$).

## 2.2. Information spaces and filters

We can reformulate the notion of a *history information space* introduced by LaValle (2006) as follows. In this context, $X$ is an external state space that characterizes the robot's configuration, velocity, and environment, $U$ is an action space, $f$ is a state transition mapping that produces a next state from a current state and action, $h$ is a sensor mapping that maps states to observations, and $Y$ is a sensor observation space. As in LaValle (2006), for each $x \in X$, let $\Psi(x)$ be a finite set of "nature sensing actions" and for each $x \in X$ and $u \in U$ let $\Theta(x, u)$ be a finite set of "nature actions." Let $X_\Psi = \{(x, \psi) \mid \psi \in \Psi(x)\}$ and let $h: X_\Psi \to Y$ be a "sensor mapping" where $Y$ is a set called the "observation space." Let $X_\Theta = \{(x, u, \theta) \mid \theta \in \Theta(x, u)\}$ and let $f: X_\Theta \to X$ be the "transition function." The following definition is an adaptation from LaValle (2006).

**Definition 2.5.** A *valid history I-state for $X, \Psi, \Theta, f$* is a sequence $(u_0, y_0, \ldots, u_{k-1}, y_{k-1})$ of length $2k$ for which there exist $\bar{x} = (x_0, \ldots, x_{k-1})$, $\bar{\psi} = (\psi_0, \ldots, \psi_{k-1})$ and $\bar{\theta} = (\theta_0, \ldots, \theta_{k-2})$ such that for all $i < k$ we have

1. $\theta_i \in \Theta(x_i, u_i)$,
2. if $i < k - 1$, then $x_{i+1} = f(x_i, u_i, \theta_i)$,
3. $\psi_i \in \Psi(x_i)$,
4. $y_i = h(x_i, \psi_i)$.

In this case we say that $(u_0, y_0, \ldots, u_{k-1}, y_{k-1})$ is *witnessed* by $\bar{x}$, $\bar{\psi}$ and $\bar{\theta}$.

Now let $\mathcal{I}$ be the set of all valid history I-states for $X, \Psi, \Theta, f$. For all $k \in \mathbb{N}$, all $\bar{x} \in X^{k-1}$, all $\bar{\psi} = (\psi_0, \ldots, \psi_{k-1})$ and all $\bar{\theta} = (\theta_0, \ldots, \theta_{k-2})$, let $\mathcal{I}^k(\bar{x}, \bar{\psi}, \bar{\theta})$ be the set of all valid paths $(u_0, y_0, \ldots, u_{k-1}, y_{k-1})$ witnessed by $\bar{x}$, $\bar{\psi}$, and $\bar{\theta}$. Now let $T \subseteq \mathcal{I} \times (U \times Y) \times \mathcal{I}$ be defined by

$$T = \Big\{(\eta, (u, y), \eta') \mid \text{ there exist } k \in \mathbb{N},$$
$$\bar{x} = (x_0, \ldots, x_{k-1}), \bar{\psi} = (\psi_0, \ldots, \psi_{k-1}),$$
$$\bar{\theta} = (\theta_0, \ldots, \theta_{k-2}), \theta \in \Theta(x_{k-1}, u) \text{ and } \psi \in \Psi(f(x_{k-1}, u, \theta))$$
$$\text{such that}$$
$$\eta \in \mathcal{I}^k(\bar{x}, \bar{\psi}, \bar{\theta}) \wedge \eta' \in \mathcal{I}^{k+1}(\bar{x}', \bar{\psi}', \bar{\theta}'),$$
$$\text{where } \bar{x}' = \bar{x}^\frown(f(x_{k-1}, u, \theta)),\ \bar{\psi}' = \psi^\frown(\psi),\ \text{and } \bar{\theta}' = \theta^\frown(\theta)\Big\}.$$

Here, $x^\frown y$ is the concatenation of sequences $x$ and $y$. Then $(\mathcal{I}, U \times Y, T)$ is the *history I-space transition system*.

Suppose for each $x, y \in X$ there is at most one $u \in U$ with $x \xrightarrow{u} y$. Let

$$E_T = \{(x, y) \in X^2 \mid \exists u \in U(x \xrightarrow{u} y)\},$$

and let $l: E_T \to U$ be defined so that $l((x, y))$ is the unique $u$ such that $x \xrightarrow{u} y$. Then $(X, E_T, l, x_0)$ with $x_0 \in X$ is a passive I-state graph as in O'Kane and Shell (2017, Def 1).

The following definition is more of a notational than mathematical value.

**Definition 2.6.** Let $\mathcal{X} = (X, U, T)$ be a transition system. If for all $(x, u) \in X \times U$ there is a unique $y \in X$ with $(x, u, y) \in T$, then we denote the function $(x, u) \mapsto y$ by $\tau$, and write $(X, U, \tau)$ instead of $(X, U, T)$. In this case we call $\mathcal{X}$ an *automaton*. Note that usually in computer science literature an automaton is finite and also has an initial state and a set of accepting states, but we do not have those in our definition.

For automata we also use the notation $x * u = \tau(x, u)$ and if $\bar{u} = (u_0, \ldots, u_{k-1})$, then $x * \bar{u}$ is defined by induction for $k > 1$ as follows: $x * (u_0, \ldots, u_{k-1}) = (x * (u_0, \ldots, u_{k-2})) * u_{k-1}$.

**Examples 2.7.** Automata and transition systems can model agent-environment and related dynamics.

1. If $(X, \cdot)$ is a group, $U \subseteq X$ is a set of generators, and $\tau(x, u) = x \cdot u$, then $(X, U, \tau)$ is an automaton. For example, consider the situation in which $X = \mathbb{Z} \times \mathbb{Z}$ and $U = \{a, b, a^{-1}, b^{-1}\}$ in which $a = (1, 0)$ and $b = (0, 1)$. Thus, $X$ is presented with generators $a$, $b$, and relation $a \cdot b = b \cdot a$. This models an agent moving without rotation in an infinite 2D-grid and the agent can move left, right, up and down. There are no obstacles. The standard Cayley graph is equivalent to the graph based representation of the automaton.
2. Let $U^*$ be the set of all finite sequences ("strings") of elements of $U$. If $\bar{u} = (u_0, \ldots, u_{k-1}) \in U^*$ and $u_k \in U$, we denote by $\bar{u}^\frown u_k$ the *concatenation* $(u_0, \ldots, u_{k-1}, u_k)$. If $\bar{u}_0, \bar{u}_1 \in U^*$, then $\bar{u}_0{}^\frown \bar{u}_1$ is similarly the concatenation of two strings. The operation of concatenation turns $U^*$ into a monoid. Suppose $\tau : X \times U^* \to X$ is an action of the monoid $U^*$ on $X$ meaning that it satisfies $\tau(\tau(x, \bar{u}), \bar{u}') = \tau(x, \bar{u}^\frown \bar{u}')$ and $\tau(x, \varnothing) = x$. Then the automaton $(X, U, \tau)$ is a discrete-time control system. A sequence of *controls* $\bar{u} = (u_0, \ldots, u_{k-1})$ produces a unique *trajectory* $(x_0, \ldots, x_k)$, given the initial state $x_0$ by induction: $x_{i+1} = \tau(x_i, u_i)$ for all $i < k$.
3. Consider an automaton $(X, U, \tau)$ in which $U$ is a group, and $\tau$ is a group action of $U$ on $X$. In some situations it can be natural to consider the set of motor-outputs of an agent to be a group: the neutral element is no motor-output at all, every motor-output has an "inverse" for which the effect is the opposite, or negating (say, moving right as opposed to moving left), the composition of movements is many movements applied consecutively. The action $\tau$ of $U$ on $X$ is then the realization of those motor-outputs in the environment. In realistic scenarios, however, this is not a good way to model the sensorimotor interaction because of the following reason. Suppose the agent has actions "left" and "right," but it is standing next to an obstacle on its left. Then moving "left" will result in staying still (because of the obstacle), but moving "right" will result in actually moving right, if there is no obstacle at the right of the agent. In this situation the sequence "left-right" results in a different position of the agent than the sequence "right-left," so if "left" and "right" are each other's inverses in $G$, then the axioms of group action are violated.
4. Note that if $T = \varnothing$, then $(X, U, T)$ is a transition system.
5. Let $X = \{0, 1\}^*$ as in (2), $U = \{0\}$, and $(x, 0, y) \in T$ if and only if $|y| = |x| + 1$, then $(X, U, T)$ is a transition system, where $|x|$ is the length of the string $x$.
6. If $(X, U, T)$ is a transition system and $E \subseteq X$ an equivalence relation, then $(X/E, U, T/E)$ is a transition system, where $X/E = \{[x]_E \mid x \in X\}$ and $T/E = \{([x]_E, u, [y]_E) \mid (x, u, y) \in T\}$, and $/$ denotes a quotient space; see Definition 2.33.

## 2.3. Sensorimotor systems

Next, we will define a *sensorimotor system*, which is a special case of a transition system. Following the tenet (EA1) that "environment is inseparable from the body which is inseparable from the brain," our sensorimotor systems can model any part of the environment-body-brain coupling. The model that describes the environment differs from the one that describes the agent merely in the type of structure it possesess, but not in an essential mathematical way.

SM-systems can be thought of as a partial specification of (some part of) the brain-body-environment coupling. Physicalist determinism demands that under full specification[2] we are left with a deterministic system. A specification is partial when it leaves room for unknowns in some, or all, parts of the system.

**Definition 2.8.** A *sensorimotor system* (or *SM-system*) is a transition system $(X, U, T)$ where $U = S \times M$ for some sets $S$ and $M$, which we call in this context the *sensory set* and the *motor set*, *respectively*.

The interpretation is that if $x \xrightarrow{(s,m)} y$, then $s$ is the sensation that either occurs at $x$, or along the transition to the next state $y$, and $m$ the motor action which leads to the transition. We will show later how SM-systems can be connected together (Definition 2.22) to form coupled systems. Sometimes an SM-system is modeling a brain-body totality, and other times it is modeling body-environment totality. A coupling between these two will model the brain-body-environment totality. This is a flexible framework which enables enactivist-style analysis. We do not assume that the agent "knows" the effect of a given $m \in M$ or that the "meaning" of a given $s \in S$. The sets $S$ and $M$ are purely mathematical sets denoting the interface between the agent and the environment from the third person perspective.

In fact, the sensory and motor components can be decoupled which might be more natural from the mathematics' point of view in some cases. The following shows that we can look at it both ways.

2 This means a full specification of the environment, the agent's body, its brain, their coupling, as well as the initial states.

**Definition 2.9.** An *asynchronous SM-system* is a transition system $(X, U, T)$ such that there exist partitions $U = S \cup M$ and $X = X_s \cup X_m$ such that for all $(x, u, y) \in T$ we have

1. if $x \in X_s$, then $u \in S$,
2. if $x \in X_m$, then $u \in M$, and
3. $x \in X_m \iff y \in X_s$.

Thus, the state space of a sequential SM-system contains separate *sensory states* and *motor states*.

**Definition 2.10.** Suppose $E$ is an equivalence relation on a set $X$. We say that a map $f : X \to X$ is *E-preserving* if for all $x, y \in X$, we have $xEy \iff f(x)Ef(y)$.

There is a natural correspondence between SM-systems and their asynchronous counterpart:

**Theorem 2.11.** *Let* SM *and* aSM *be the classes of SM-systems and asynchronous SM-systems, respectively. There are functions* $F$: SM $\to$ aSM *and* $G$: aSM $\to$ SM *such that*

1. *F and G are isomorphism and bisimulation preserving,*
2. *restricted to finite systems, F and G are polynomial-time computable, and restricted to the infinite ones they are Borel-functions in the sense of classical descriptive set theory (Kechris, 1994).*

*Proof:* See Appendix B. □

Another type of a system, which is in a similar way equivalent to a special case of an SM-system, is a state-labeled transition system which we will introduce next, and prove a similar result, Lemma 2.19.

## 2.4. Quasifilters and quasipolicies

The amount of information specified in a given SM-system depends on which part of the brain-body-environment system we are modeling. At one extreme, we specify the environment's dynamics down to the small detail and leave the brain's dynamics completely unspecified. In this case the SM-system will have only one sensation corresponding to each state and the transition to the next state will be completely determined by knowing the motor action. This is, in a sense, the environment's perspective. At the other extreme, we specify the brain completely, but leave the environment unspecified. We "don't know" which sensation comes next, but we "know" which motor actions are we going to apply. This is in a sense the perspective of the agent. The first extreme case is the perspective often taken in robotics and other engineering fields when either specifying a planning problem (Ghallab et al., 2004; Choset et al., 2005; O'Kane and LaValle, 2008), or designing a filter (Hager, 1990; Thrun et al., 2005; LaValle, 2012; Särkkä, 2013) (also known as sensor fusion). This is why we call SM-systems of that sort *quasifilters* (Definition 2.12). The other extreme is the perspective of a policy. The policy depends on sensory input, but the motor actions are determined (by the policy). This is why we call the SM-systems of the latter sort *quasipolicy*. The "quasi-" prefix is used because both are weaker and more general notions than those that appear in the literature; see Remarks 2.20 and 2.21.

Another way to look at this is the dichotomy between virtual reality (VR), and robotics. In virtual reality, scientists are designing the (virtual) environment for an agent whereas in robotics they are typically designing an agent for an environment. In the former case the agent is partially specified: the type of embodiment is known ($S$ and $M$ are known) and some types of patterns of embodiment are known (eye-hand coordination). However, the specific actions to be taken by the agents are left unspecified. The job of the designer is to specify the environment down to the smallest detail, so that every sequence of motor actions of the agent yields targeted sensory feedback. The VR-designer is designing a quasifilter constrained by the partial knowledge of the agent's embodiment and internal dynamics. The case for the robot designer is the opposite. She has a partial specification of the robot's intended environment and usually works with a complete specification of the robot's mechanics. She is designing a quasipolicy. For VR-designers the agent is a black box; for roboticists the agent is a white box (Suomalainen et al., 2020) (unless the task is to reverse engineer an unknown robot design). For the environment, the roles are reversed. A similar dichotomy can be seen between biology (in which the agent is a black box) and robotics (in which it usually is a white box).

All the definitions in this section are new.

**Definition 2.12.** Suppose that $(X, S \times M, T)$ is an SM-system with the property that for all $x_1 \in X$ there exists $s_{x_1} \in S$ such that for all $x_2 \in X$ and all $(s, m) \in S \times M$ we have that $x_1 \xrightarrow{(s,m)} x_2$ implies $s = s_{x_1}$. Then, $(X, S \times M, T)$ is a *quasifilter*.

In a quasifilter the sensory part of the outgoing edge is unique. The dual notion (quasipolicy) is when the motor part is unique:

**Definition 2.13.** Suppose that $(X, S \times M, T)$ is an SM-system with the property that for all $x \in X$ there exists $m_x \in M$ such that for all $y \in X$ and all $(s, m) \in S \times M$ we have that $x \xrightarrow{(s,m)} y$ implies $m = m_x$. Then, $(X, S \times M, T)$ is a *quasipolicy*.

Before explaining the connections between quasifilter and a filter and quasipolicy and a policy, let us define projections of the sensorimotor transition relation to "motor" and to "sensory":

**Definition 2.14.** Given an SM-system $(X, S \times M, T)$, let

$$T_M = \{(x, m, y) \in X \times M \times X \mid \exists s \in S (x, (s, m), y) \in T\}$$
$$T_S = \{(x, s, y) \in X \times S \times X \mid \exists m \in M (x, (s, m), y) \in T\}.$$

These are called the *motor* and the *sensory projections*, respectively of the sensorimotor transition relation. They are

also called the *motor transition relation* and the *sensory transition relation*, respectively. The corresponding transition systems $(X, M, T_M)$ and $(X, S, T_S)$ are called the *motor* and the *sensory projection systems*.

**Definition 2.15.** Given a transition system $(X, U, T)$, and $x \in X$, let $O_T(x) \subseteq U$ be defined as the set $O_T(x) = \{u \in U \mid (\exists y \in X)(x \xrightarrow{u} y)\}$. Combining this notation with the one introduced in Example 2.4(2), given $x, y \in X$, we have

$$O_T(x) = \bigcup_{y \in X} \tilde{T}(x, y).$$

For a transition relation $T \subseteq X \times (S \times M) \times X$, define its *transpose* by $T^t \subseteq X \times (S \times M) \times X$ such that $T^t = \{(x, (m, s), y) \mid (x, (s, m), y) \in T\}$. Note that $(T^t)^t = T$. For a subset of a Cartesian product $A \subseteq S \times M$, let $A_1$ be the projection to the first coordinate $A_1 = \{s \in S \mid (\exists m \in M)((s, m) \in A)\}$ and $A_2$ the projection to the second one: $A_2 = \{m \in M \mid (\exists s \in S)((s, m) \in A)\}$.

Mathematically coupling of two transition systems is symmetric [see Theorem 2.24(3)], but from the cognitive perspective there is (usually) an asymmetry between the agent and the environment (which can be evident from some specific properties of the agent and of the environment). Because of the partial symmetry, many properties of an agent can dually be held by the environment and vice versa. The following proposition highlights the duality between quasipolicy and quasifilters: reversing the roles of the environment and the agent.

**Proposition 2.16.** *For an SM-system $\mathcal{X} = (X, S \times M, T)$ the following are equivalent:*

1. *$\mathcal{X}$ is a quasifilter,*
2. *$\mathcal{X}^t = (X, S \times M, T^t)$ is a quasipolicy,*
3. *$O_{T_S} = (O_T(x))_2 = (O_{T^t}(x))_1$ is a singleton for each $x \in X$.*

*Similarly, $\mathcal{X}$ is a quasipolicy if and only if $O_{T_M}(x) = (O_T(X))_1$ is a singleton for each $x \in X$.*

*Proof:* A straightforward consequence of all the definitions. □

## 2.5. State-relabeled transition systems

It will become convenient in the coming framework to assign labels to the states. The elements $x$ of the state space $X$ are already named; thus, our labeling can be more properly considered as a *relabeling via* a function $h : X \to L$, in which $L$ is an arbitrary set of *labels*. This allows partitions to be naturally induced over $X$ by the preimages of $h$. Intuitively, this will allow the state space $X$ to be characterized at different levels of "resolution" or "granularity." Thus, we have the following definition:

**Definition 2.17.** A *state-relabeled transition system* (or simply *labeled transition system*) is a quintuple $(X, U, T, h, L)$ in which $h : X \to L$ is a labeling function and $(X, U, T)$ is a transition system.

We think of *state-relabeled* to be a more descriptive term, but we shorten it in the remainder of this paper to being simply *labeled*.

*Remark* 2.18. In an analogy to Definition 2.6, a labeled transition system is a *labeled automaton*, if $T$ happens to be a function; in other words, for all $(x, u) \in X \times U$ there is a unique $y \in X$ with $(x, u, y) \in T$. In this case we may denote this function by $\tau : (x, u) \mapsto y$ and work with the labeled automaton $(X, U, \tau, h, L)$. For example, the temporal filter in Section 2.1 is a labeled automaton.

The isomorphism and bisimulation relations are defined similary as for transition systems, but in a label-preserving way.

One intended application of a labeled transition system $(X, U, T, h, L)$ is that $h$ is a sensor mapping, $L$ is a set of sensor observations, and $U$ is a set of actions. Thus, actions $u \in U$ allow the agent to transition between states in $X$ while $h$ tells us what the agent senses in each state. We intend to show that this can be seen as a special case of an SM-system by proving a theorem similar to Theorem 2.11, but stronger, namely these corresponces preserve isomorphism:

**Lemma 2.19.** *Let $\mathcal{F}$ be the class of quasifilters, $\mathcal{P}$ the class of quasipolicies, and $\mathcal{L}$ the class of labeled systems. Then there are one-to-one maps*

$$\mathrm{LTS}_P : \mathcal{P} \to \mathcal{L} \text{ and } \mathrm{LTS}_F : \mathcal{F} \to \mathcal{L}$$

*such that*

1. *$\mathrm{LTS}_P$ and $\mathrm{LTS}_F$ are isomorphism and bisimulation preserving,*
2. *restricted to finite systems, $\mathrm{LTS}_P$ and $\mathrm{LTS}_F$ are polynomial-time computable, and restricted to the infinite ones they are Borel-functions in the sense of classical descriptive set theory.*

*Proof:* See Appendix B □

*Remark* 2.20. Let $\mathcal{X} = (X, S \times M, T)$ be a quasifilter and $\mathcal{X}' = \mathrm{LTS}_F(\mathcal{X}) = (X, M, T_M, h, S)$ as in Lemma 2.19. Suppose further that for each $x, y \in X$ there is at most one $u \in U$ with $x \xrightarrow{u} y$. Let

$$E_T = \{(x, y) \in X^2 \mid \exists u \in U(x \xrightarrow{u} y)\},$$

Then $(X, M, E_T, x_0)$ coincides with the definition of a filter (O'Kane and Shell, 2017, Def 3). If it is also an automaton, meaning that above we replace "at most one" by "exactly one," then every sequence of motor actions $(m_0, \ldots, m_{k-1})$ determines a unique resulting state $x_{k-1} \in X$. This is analogous, and can be proved in the same way, as the fact that each sequence of sensory data determines a unique resulting state in Remark 2.21 below.

*Remark* 2.21. Usually, a *policy* is a function which describes how an agent chooses actions based on its own past experience. Thus, if $M$ is the set of motor commands and $S$ is the set of sensations,

a policy is a function $\pi : S^* \to M$ where $S^*$ is the set of finite sequences of sensory "histories"; see for example (LaValle, 2006). Now, suppose that an SM-system $\mathcal{X} = (X, S \times M, T)$ is a quasipolicy in the sense of Definition 2.13 and let $x \mapsto m_x$ be as in that Definition. Assume further that $\mathcal{X}$ is an automaton (Section 2.1) and let $\tau : X \times (S \times M) \to X$ be the corresponding transition function so that for all $x \in X$ and $(s, m) \in S \times M$ we have $(x, (s, m), \tau(x, (s, m))) \in T$. Let $x_0 \in X$ be an initial state. We will show how the pair $(\mathcal{X}, x_0)$ defines a function $\pi : S^* \to M$ in a natural way. Let $\bar{s} = (s_0, \ldots, s_{k-1}) \in S^k$ be a sequence of sensory data. If $k = 0$, and so $\bar{s} = () = \varnothing$, let $\pi(\bar{s}) = m_{x_0}$. If $k > 0$, assume that $\pi(s_0, \ldots, s_{k-2})$ and $x_{k-1}$ are both defined (induction hypothesis). Then let $x_k = \tau(x_{k-1}, (m_{x_{k-1}}, s_{k-1}))$ and $\pi(s_0, \ldots, s_{k-2}, s_{k-1}) = m_{x_k}$. The idea is that because of the uniqueness of $m_x$, a sequence of sensory data determines (given an initial state) a unique path through the automaton $\mathcal{X}$.

## 2.6. Couplings of transition systems

The central concept of this work pertaining to all principles (EA1)–(EA5) is the coupling of SM-systems. We define coupling, however, for general transition systems with the understanding that our most interesting applications will be for SM-systems where $U_0 = U_1 = S \times M$. The idea is that in every transition there is a sensory component and a motor component. The set $S$ could be thought of as all possible events that trigger afferent nervous signals, or their combinations. The elements of $M$ are those events that are triggered by efferent nervous signals. This is an abstract space and in transitioning from one state to another some subset of $S \times M$ is "active." If we know little of what kind of sensory data the agent receives during the transition, then that transition will occupy a subset of $S \times M$ whose projection to the $S$-coordinate is large. If, on the other hand we know a lot, and can specify the exact sensory data, then the projection to the $S$-coordinate is small. Vice versa, if we do not know which motor actions lead from one state to another, then the projection of the corresponding subset to the $M$-coordinate is large etc. This was made more precise in Section 2.4. The fact that the transition consists of pairs $(s, m)$ where $s$ is a sensory input and $m$ is a motor command does not mean that the agent is equipped with the semantics of what $m$ "means," or what it "does" in the world. The effect of $m$ is "computed" by the environment and the agent only receives the next "$s$" as the feedback. It might have been more intuitive, but more cumbersome to make this definition in terms of functions that map events of the environment to sensory stimuli and internal events of the nervous system to motor actions, and further functions that map the motor actions to the actual events in the environment, etc., but from the point of view of essential mathematical structure these extra identifications wouldn't add anything qualitatively new.

**Definition 2.22.** Let $\mathcal{X}_0 = (X_0, U_0, T_0)$ and $\mathcal{X}_1 = (X_1, U_1, T_1)$ be two transition systems. The *coupled* system $\mathcal{X}_0 * \mathcal{X}_1$ is the transition system $(X, U, T)$ defined as follows: $X = X_0 \times X_1$, $U = U_0 \cap U_1$, and

$$T = T_0 * T_1 = \{((x_0, x_1), u, (y_0, y_1)) \mid (x_0, u, y_0) \in T_0 \wedge (x_1, u, y_1) \in T_1\}.$$

Equivalently, for all $((x_0, x_1), (y_0, y_1)) \in (X_0 \times X_1)^2$ we have

$$\tilde{T}((x_0, y_0), (x_1, y_1)) = \tilde{T}_0(x_0, x_1) \cap \tilde{T}_1(y_0, y_1)$$

(recall the $\tilde{T}$ notation from Example 2.4(2)).

**Example 2.23.** A simple example of coupling is illustrated in Figure 2.

Mathematically the coupling is a product of sorts. If we think of one transition system as "the environment" and the other as "the agent," then the coupling tells us about all possible ways in which the agent can engage with the environment. The fact that the state space of the coupled system is the product of the state spaces of the two initial systems reflects the fact that the coupled system includes information of "what would happen" if the environment was in any given state while the agent is in any given ("internal") state.

We immediately prove the first theorem concerning coupling:

**Theorem 2.24.** *Suppose that $\mathcal{X}_i = (X_i, U_i, T_i)$ and $\mathcal{X}'_i = (X'_i, U'_i, T'_i)$ for $i \in \{0, 1\}$ are four SM-systems. Then the following hold:*

1. *If $\mathcal{X}_i \cong \mathcal{X}'_i$ for $i \in \{0, 1\}$, then $\mathcal{X}_0 * \mathcal{X}_1 \cong \mathcal{X}'_0 * \mathcal{X}'_1$.*
2. *If $\mathcal{X}_i \sim \mathcal{X}'_i$ for $i \in \{0, 1\}$, then $\mathcal{X}_0 * \mathcal{X}_1 \sim \mathcal{X}'_0 * \mathcal{X}'_1$.*
3. *$\mathcal{X}_0 * \mathcal{X}_1 \cong \mathcal{X}_1 * \mathcal{X}_0$.*

*Proof:* See Appendix B □

Coupling provides an interesting way to compare SM-systems from the "point of view" of other SM-systems. For example, given an SM-system $\mathcal{E}$ one can define an equivalence relation on SM-systems by saying that $\mathcal{I} \sim_E \mathcal{I}'$, if $\mathcal{E} * \mathcal{I} = \mathcal{E} * \mathcal{I}'$. If $\mathcal{E}$ is the "environment" and $\mathcal{I}, \mathcal{I}'$ are "agents," this is saying that the agents perform identically in this particular environment. Or vice versa, for a fixed $\mathcal{I}$, the relation $\mathcal{E} * \mathcal{I} = \mathcal{E}' * \mathcal{I}$ means that the environments are indistinguishable by the agent $\mathcal{I}$.

*Remark* 2.25. In the definition of coupling we see that the two SM-systems constrain each other. This is seen from the fact that in the definition we take intersections. For example, when an agent is coupled to an environment, it chooses certain actions from a large range of possibilities. In this way the agent structures its own world through the coupling (EA3). To make this notion further connect to enactivist paradigm, we invoke the dynamical systems

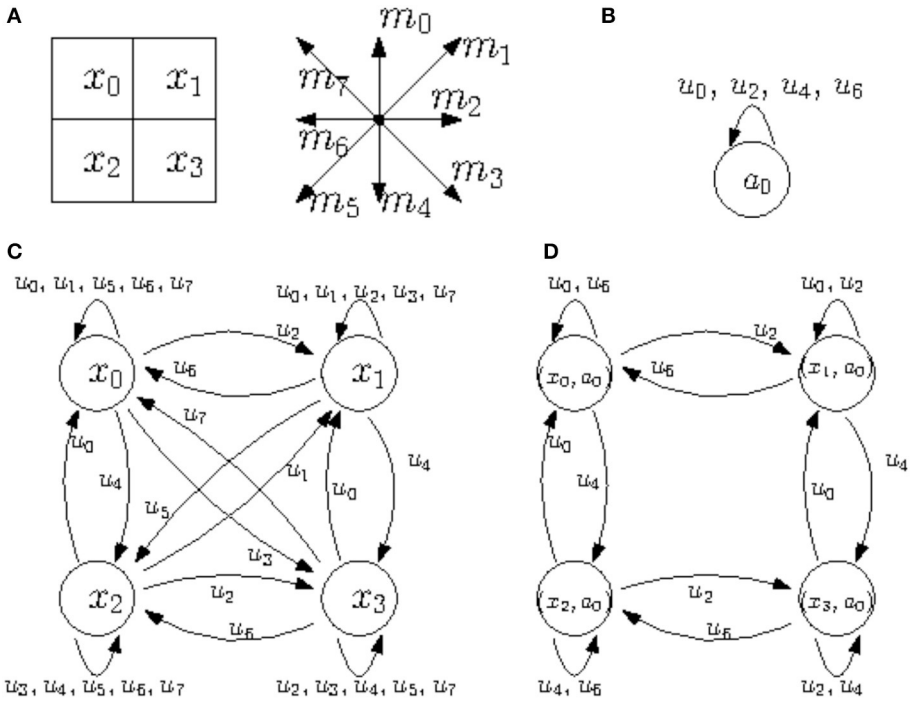



FIGURE 2
**(A)** States and actions for the transition system $\mathcal{X}_0$ that describes a 2-by-2 grid. 8 actions populating the set $M = \{m_0, \ldots, m_7\}$ correspond to a move (to a neighbor cell if possible) either sideways or diagonally. Suppose $S$ is a singleton such that $S = \{s\}$. Then, in the following, $u_i$ corresponds to the transition $u_i = (m_i, s)$ for $i = 1, \ldots, 7$. **(B)** Transition system $\mathcal{X}_1$. **(C)** Transition system $\mathcal{X}_0$. **(D)** The coupled system $\mathcal{X}_0 * \mathcal{X}_1$.

approach to cognition (Tschacher and Dauwalder, 2003). An *attractor* in a transition system $\mathcal{X} = (X, U, T)$ is a set $A \subseteq X$ with the property that for all infinite sequences

$$x_0 \overset{u_0}{\to} x_1 \overset{u_1}{\to} \cdots x_{k-1} \overset{u_{k-1}}{\to} x_k \overset{u_k}{\to} \cdots$$

there are infinitely many indices $n$ such that $x_n \in A$. There could be other possible definitions, such as "for all large enough $n$, $x_n \in A$". For the present illustration purposes it is, however, irrelevant. It could be the case that $A \subseteq X$ is not an attractor of $\mathcal{X}$, but after coupling with $\mathcal{X}' = (X', U', T')$, $A \times X'$ may be an attractor of $\mathcal{X} * \mathcal{X}'$. Thus, if $\mathcal{X}$ is the environment and $\mathcal{X}'$ is the agent and $A$ is a set of desirable environmental states, then we may say that the agent is well attuned to $\mathcal{X}$, if $A$ was not initially an attractor, but in $\mathcal{X} * \mathcal{X}'$, then $A \times X'$ becomes one. It could also be that the agent needs to arrive to $A$ while being in a certain type of an internal state $B \subseteq X'$, for example, if $A$ is "food" and $B$ is "hungry". Then it is not important that $A \times X'$ is an attractor, but it is imperative that $A \times B$ is one.

## 2.7. Unconstrained and fully constrained SM-systems

As we mentioned before, the information specified in an SM-system depends on which part of the brain-body-environment system we are modeling. In the extreme case we do not specify *anything*, except for the very minimal information. Consider a body of a robot for which the set of possible actions (or motor commands) is $M$ and the set of possible sensor observations is $S$. Suppose that is all we know about the robot. We do not know what kind of environment it is in and we do not know what kind of "brain" (a processor or an algorithm) it is equipped with. Thus, we do not know of any constraints the robot may have in sensing or moving. We then model this robot as an *unconstrained* SM-system:

**Definition 2.26.** An SM-system $(X, S \times M, T)$ is called *unconstrained* iff for all $x \in X$, we have $O_T(x) = S \times M$; recall Definition 2.15.

Unconstrained systems have the role of a neutral element with respect to coupling (Proposition 2.29). We now show that given all unconstrained SM-systems with shared $M$ and $S$ are mutually bisimulation equivalent:

**Proposition 2.27.** *Suppose that* $\mathcal{X} = (X, S \times M, T)$ *and* $\mathcal{X}' = (X', S \times M, T')$ *are unconstrained systems. Then* $\mathcal{X} \sim \mathcal{X}'$.

*Proof:* See Appendix B □

There are many intuitions behind the above. An unconstrained system is one where anything could happen: the agent might perform any actions in any order and the environment could provide the agent with any sensory data. Such a world is reminiscent of *white noise*. Such a system is only interesting from an abstract mathematical perspective, it is in some sense "maximal". The content of Proposition 2.27 is that such systems are indistinguishable from each other. An unconstrained system has a similar role with respect to all SM-systems as the free group has to other groups, although we haven't made this universality claim precise in the present paper. Intuitively it means that every possible agent-environment combination can be found as a subsystem (or possibly a quotient) of the unconstrained one. The term "unconstrained" refers in particular to that when coupled to other systems, this system doesn't constrain them, so it acts in the same way as 0 in arithmetic addition (Proposition 2.29). The opposite is the fully constrained system (Definition 2.31, Proposition 2.32). In that case, the intuition is the opposite: in environments where nothing happens and actions do not have any effects, any agent is as good as any other and vice versa: agents that don't do anything are equivalent.

**Corollary 2.28.** *The SM-system* $\varepsilon = (\{0\}, \{0\} \times (S \times M) \times \{0\}\})$ *is the unique, up to bisimulation, unconstrained system.*

**Proposition 2.29.** *Let* $\varepsilon$ *be as in Corollary 2.28 and let* $\mathcal{X} = (X, S \times M, T)$ *be any SM-system. Then* $\mathcal{X} * \varepsilon \cong \mathcal{X}$.

**Corollary 2.30.** *If* $\mathcal{X}$ *and* $\mathcal{X}'$ *are SM-systems and* $\mathcal{X}'$ *is unconstrained, then* $\mathcal{X} * \mathcal{X}' \sim \mathcal{X}$.

*Proof:* By Corollary 2.28 $\mathcal{X}' \sim \varepsilon$, So by Theorem 2.24 we have $\mathcal{X} * \mathcal{X}' \sim \mathcal{X} * \varepsilon$. However, by Proposition 2.29, $\mathcal{X} * \varepsilon \sim \mathcal{X}$; thus, $\mathcal{X} * \mathcal{X}' \sim \mathcal{X}$. □

The opposite of an unconstrained system is a fully constrained one:

**Definition 2.31.** An SM-system $(X, S \times M, T)$ is *fully constrained* iff $T = \varnothing$.

**Proposition 2.32.** *Dually to the propositions above, we have that (1) all fully constrained systems are bisimulation equivalent to each other, (2) the simplest example being* $\lambda = (\{0\}, S \times M, \varnothing)$, *and (3) if* $\mathcal{X} = (X, S \times M, T)$ *is another SM-system, then* $\mathcal{X} * \lambda \sim \lambda$.

All transition systems are in some sense between the fully constrained and the unconstrained, these being the two theoretical extremes.

## 2.8. Quotients of transition systems

When considering labelings and their induced equivalence relations, it will be convenient to develop a notion of quotient systems, analogous to quotient spaces in topology. Suppose $\mathcal{X} = (X, U, T)$ is a transition system and $E$ is an equivalence relation on $X$. We can then form a new transition system, called the *quotient* of $\mathcal{X}$ by $E$ in which the new states are $E$-equivalence classes and the transition relation is modified accordingly.

The following definition of a quotient is standard in Kripke model theory, especially bisimulation theory:

**Definition 2.33.** Suppose $\mathcal{X} = (X, U, T)$ and $E$ are as above. Let $X/E = \{[x]_E \mid x \in X\}$, in which each $[x]_E$ is an equivalence class of states $x$ under relation $E$, and $T/E = \{([x]_E, u, [y]_E) \mid (x, u, y) \in T\}$. Then $\mathcal{X}/E = (X/E, U, T/E)$ is the *quotient* of $(X, U, T)$ by $E$.

The following definition is inspired by the idea of sensory pre-images, see LaValle (2019), but is also needed for technical reasons.

**Definition 2.34.** Given any function $h\colon X \to L$, denote by $E^h$ the inverse-image equivalence: $E^h = \{(x, y) \in X^2 \mid h(x) = h(y)\}$. We will denote the equivalence classes of $E^h$ by $[x]_h$ instead of $[x]_{E^h}$ if no confusion is possible.

The equivalence relation $E^h$ partitions $X$ according to the preimages of $h$, as considered in the sensor lattice theory of LaValle (2019). The partition of $X$ induced by $h$ directly yields an quotient transition system by applying the previous two definitions:

**Definition 2.35.** Let $\mathcal{X} = (X, U, T)$ be a transition system and $h\colon X \to L$ be any mapping. Then define $\mathcal{X}/h$ to be $\mathcal{X}/E^h$ where we combine Definitions 2.34 and 2.33.

**Proposition 2.36.** *If* $h$ *is one-to-one, then* $\mathcal{X}/h \cong \mathcal{X}$.

*Proof:* $h$ is one-to-one if and only if $E^h$ is equality, in which case it is straightforward to verify that the function $x \mapsto [x]_{E^h}$ is an isomorphism. □

For $h\colon X \to L$, the transition system $(X/h, U, T/h)$ is essentially a new state space over the preimages of $h$. In this case $\mathcal{X}/h$ is called the *derived information space* (as used in LaValle, 2006). More precisely:

**Proposition 2.37.** *Let* $L' = \mathrm{ran}(h) \subseteq L$. *Define*

$$T' = \{(l, u, l') \in L' \times U \times L' \mid (h^{-1}(l), u, h^{-1}(l')) \in T/h\}$$
$$= \{(h(x), u, h(y)) \mid (x, u, y) \in T\}.$$

*Then* $(X/h, U, T/h)$ *is isomorphic to* $(L', U, T')$ *via the isomorphism* $f\colon [x]_{E^h} \mapsto h(x)$.

*Proof:* See Appendix B □

The intuitive meaning of the quotient is the following. There is a Soviet comedy film from the 1970's where the main character ends up in an apartment in Leningrad, while he thinks that he is actually in Moscow. The apartement in Leningrad is identical to his home in Moscow and he cannot distinguish between them. He thinks for a while that he is at his home in Moscow while being in an apartment in Leningrad. Even his key from Moscow worked for the Leningrad apartment. The pun is that in Soviet times all houses were built according to the same blueprint. Now, before he realized his situation, as far as he was concerned, he *was* in Moscow. He thought he came to the same place in the evening as in the morning, while he actually didn't. The idea of the quotient captures exactly that: We identify those states that "look the same" (the label is the same) even though they are actually different states. In fact, let us look at a cognitive system on several levels of granularity: When I type on my laptop at home or in a cafeteria, my fingers experience the keyboard in (approximately) the same way. As far as my fingers (and associated motor areas) are concerned, we can identify all situations where they are pressing keys on my keyboard. On a higher level, I might be coming home after a 10 h time and experience as if I am in the same place, but we all know that the planet, on which my home is, has moved, so I actually am not in the same place, just like the main character in the movie referenced above.

## 3. Illustrative examples of SM-systems

We next illustrate how sensorimotor systems model body-environment, brain-body, and brain-body-environment couplings. Consider a body in a fully understood and specified deterministic environment. In this case the body-environment system will be modeled by a quasifilter, Definition 2.12. Instead of using the quasifilter definition, we work with a labeled transition system which, according to Proposition 2.19, is equivalent. According to the assumption of full specification, we will in fact work with labeled automata.

The body has a set $M$ of possible motor actions each of which has a deterministic influence on the body-environment dynamics. Denote the set of body-environment states by $E_0$. Whenever a motor action $m \in M$ is applied at a body-environment state $e \in E$, a new body-environment state $A(e,m) \in E$ is achieved. At each state $e \in E$ the body senses data $\sigma(e)$. Denote the set of sensations by $S$. In this way, the labeled automaton $\mathcal{E}_0 = (E, M, A, \sigma, S)$ models this body-environment system. This model is ambivalent toward the agent's internal dynamics, its strategies, policies and so on, but not ambivalent toward its embodiment and its environment's structure. In fact, it characterizes them completely.

Alternatively, consider a brain in a body, and suppose that the brain is fully understood and deterministic (for example, perhaps it is designed by us), but we do not know which environment it is in. We model this by an SM-system which is a quasipolicy. Again, by the analogous considerations as above, we work directly an equivalent labeled automaton specification. Denote the set of internal states of the brain by $I$. The agent's internal state is a function of the sensations; therefore, let $B\colon I \times S \to I$ be a function ($B$ stands for *brain*) that takes one internal state to another based on new sensory data. At each internal state, the agent produces a motor output which is an element of the set $M$; therefore, let $\mu\colon I \to M$ be a function assigning a motor output to each internal state. Now, $\mathcal{I} = (I, S, B, \mu, M)$ is a labeled transition system modeling this agent. It is ambivalent toward the type of the environment the agent is in, but it is not ambivalent toward the agent's internal dynamics, policies, strategies and so on; in fact, it determines them completely.

Now, the coupling of the environment $\mathcal{E}$ and the agent $\mathcal{A}$ is the SM-system obtained as

$$\mathrm{LTS}_F^{-1}(\mathcal{E}) * \mathrm{LTS}_P^{-1}(\mathcal{A}).$$

The sensory and motor sets $S$ and $M$ capture the interface between the brain and the environment because they characterize the body (but not the *embodiment*).

**Example 3.1.** Consider an agent that has four motor outputs, called "up" ($U$), "down" ($D$), "left" ($L$), and "right" ($R$), and there is no sensor feedback (this defines the body). In Corollary 2.28 we gave a minimal example of an unconstrained SM-system. On the other extreme one can give large examples. For instance the free monoid generated by the set $M = \{U, D, L, R\}$.

Let $X$ be the set of all possible finite strings in the four "letter" alphabet $M$, let $T = \{(x, m, y) \mid x^\frown m = y\}$. "No sensor data" is equivalent to always having the same sensor data; thus, we can assume that $S = \{s_0\}$ is a singleton and the sensor mapping $h\colon X \to S$ is constant.[3] The resulting unconstrained transition system $\mathcal{U} = (X, T, M, \sigma, S)$ can be represented by an infinite quaternary tree, shown in Figure 3A.

Suppose that this body is situated in a $2 \times 2$ grid. The body can occupy one of the four grid's squares at a time, and when it applies one of the movements, it either moves correspondingly, or, if there is a wall blocking the movement, it doesn't. This defines the body-environment system. The set of states is now $E$ and has four elements corresponding to all the possible positions of the body. The transition function $A\colon E \times M \to E$ tells where to move, and the rest is as above. The system $\mathcal{E} = (E, A, M, \sigma, S)$ is shown in Figure 3B. Let us now look at the agent. Suppose that it applies the following policy: (1) In the beginning move left; (2) if the previous move was to the left, then move right, otherwise

3 We do not mean to say that no data is always the same as some other data. We are talking here about an agent that *never* receives any data, or an agent that *always* receives the same data. Thus, it cannot rely on any "change" between having and not having any sensory input. Thus, there is no "presense in absence" paradox here.

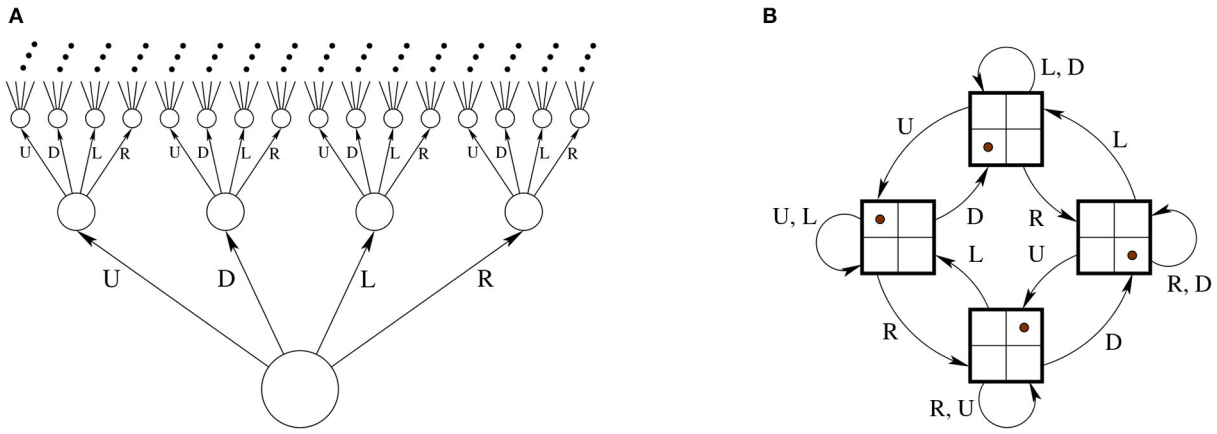



FIGURE 3
**(A)** Having motor commands and no sensory feedback leads to an infinite tree automaton. **(B)** Once the body is coupled with a 2 × 2 grid environment, a four-state automaton results.

move left. This can be modeled with a two-state automaton $\mathcal{I} = (I, S, B, \mu, M)$ where $I = \{L, R\}$, $S = \{s_0\}$, $B(L, s_0) = R$, $B(R, s_0) = L$, $\mu(L) = l$ and $\mu(R) = r$. Now, the coupling $\mathrm{LTS}_F^{-1}(\mathcal{E}) * \mathrm{LTS}_P^{-1}(\mathcal{I})$ is an automaton that realizes the policy in the environment, as shown in Figure 4A.

If the agent has a different embodiment in the same environment, then all of the automata will look different. Suppose that instead of the previous four actions, the agent has two: "rotate 90-degrees counterclockwise" ($C$),"forward one step" ($F$). Note that these are expressed in the local frame of the robot: It can either rotate relative to its current orientation, or it can move in the direction it is facing; the previous four actions were expressed as if in a global frame or the robot is incapable of rotation. Under the new embodiment, the unconstrained automaton with no sensor feedback is an infinite *binary tree*, with every node having two outgoing edges, labeled $C$ and $F$, respectively, instead of the quaternary infinite tree depicted on Figure 3A. Instead of the four-state automaton of Figure 3, the automaton describing the environment transitions is a 16 state-automaton, because the orientation of the agent can now have four different values. See Figure 4B. Finally the automaton describing the internal mechanics of the agent $\mathcal{I}$ is a quasipolicy in these two actions, and finally, the coupling corresponds essentially to taking a path in the 16-state automaton above.

Note that there is a bisimulation between $\mathcal{U}$ and $\mathcal{E}$ which reflects the fact that from the point of view of an agent they are indistinguishable. This is natural because there is no sensory data, so from the agent's viewpoint it is unknowable whether or not it is embedded in an environment. A bisimulation is given as follows: Let $y_0 \in Y$ be the top-right corner and $x_0 \in X$ the root of the tree. Define $R \subseteq X \times Y$ be the minimal set satisfying the following conditions:

1. $(x_0, y_0) \in R$.
2. If $(x, y) \in R$ and $m \in M$, then $(T(x, m), U(y, m)) \in R$.

**Example 3.2.** The 16-state automaton of Example 3.1 has four automorphisms corresponding to the rotation of the environment by 90 degrees counterclockwise. Each of those automorphisms corresponds to an auto-bisimulation. Mirroring is not an automorphism because the agent's rotating action fixes the orientation of the automaton.

**Example 3.3.** Figure 5 shows an example of how an automaton with non-trivial sensing could look. Jumping a little bit ahead, it will be seen that the labeling provided by $h$ in this figure is not sufficient (a notion introduced in Definition 4.2).

## 4. Sufficient refinements and degree of insufficiency

This section presents the concept of *sufficiency*, which will be the main glue between enactivist philosophy and mathematical understanding of cognition. In Section 4.1 we introduce the main concepts and explain its profound relevance to enactivist modeling and how it can be a precursor to the emergence of meaning from meaningless sensorimotor interactions. In Section 4.2 we introduce the notion of minimal sufficient refinements, prove a uniqueness result about them, and show how they are connected to the classical notions of bisimulation as well as derived information state spaces[4].

4 There could be an interesting relationship between this concept and the free energy principle proposed by K. Friston. A system which is attuned to its environment in a sufficient way can be interpreted by an inspector as a system that is making ] predictions about its environment.

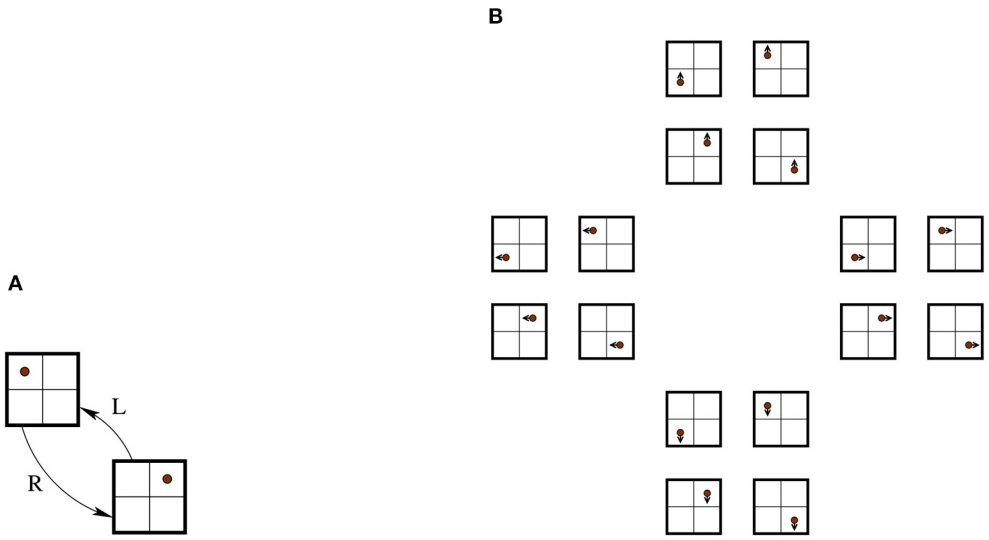



FIGURE 4
**(A)** A two-state automaton results from the realized policy. **(B)** If there are only two actions (rotate 90 degrees counterclockwise and going straight) then the second automaton has 16 states instead of four as in Figure 3B.

## 4.1. Sufficiency

The following consider the main definition of this work. It is based on the idea of sufficiency in LaValle (2006, Ch.11).

**Definition 4.1.** Let $(X, U, T)$ be a transition system and $E \subseteq X \times X$ an equivalence relation. We say that $E$ is *sufficient* or *completely sufficient*, if for all $(x, y) \in E$ and all $u \in U$, if $(x, u, x') \in T$ and $(y, u, y') \in T$, then $(x', y') \in E$.

This means that if an agent cannot distinguish between states $x$ and $y$, then there are no actions it could apply to later distinguish between them. To put it differently, if the states are indistinguishable by an instant sensory reading, then they are in fact indistinguishable even through sensorimotor interaction. This is related to the equivalence relation known as Myhill-Nerode congruence in automata theory.

The equivalence relation of indistinguishability in the context of sensorimotor interactions is at its simplest the consequence of indistinguishability by sensors. Thus, we define sufficiency for labelings or sensor mappings:

**Definition 4.2.** A labeling $h\colon X \to L$ is called *sufficient* (or *completely sufficient*) iff for all $x, y, x', y' \in X$ and all $u \in U$, the following implication holds:

$$(h(x) = h(y) \wedge (x, u, x') \in T \wedge (y, u, y') \in T) \Rightarrow h(x') = h(y').$$

**Proposition 4.3.** *If $(X, U, \tau)$ is an automaton, then $h\colon X \to L$ is sufficient if and only if for all $x, y \in X$ and all $u \in U$, we have that if $h(x) = h(y)$, then $h(\tau(x, u)) = h(\tau(y, u))$.*

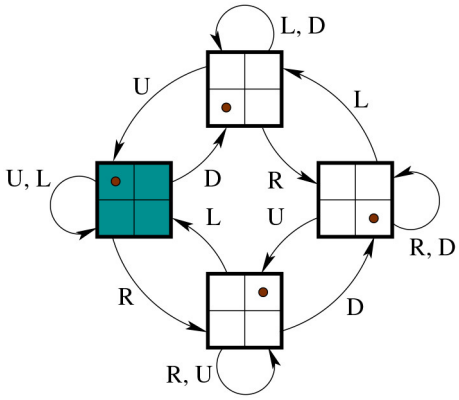



FIGURE 5
Consider the automaton $\mathcal{E}$ of Figure 3B from Example 3.1, but assume that the agent can "smell" a different scent in the top-left corner. This can be modeled by having a two-element set $S = \{0, 1\}$ instead of a singleton, and $h\colon X \to \{0, 1\}$ such that $h(x) = 0$ iff $x$ is not the top-left corner. The state with a scent is shaded.

*Proof:* Checking the definitions. □

The above proposition is saying that when the sensorimotor system is deterministic, then sufficiency is equivalent to predictability.

There is a connection with the classical notion of bisimulation in classical transition systems theory (recall Definition 2.2):

**Proposition 4.4.** *An equivalence relation on a state space of an automaton $(X, U, \tau)$ is sufficient if and only if it is an autobisimulation.*

*Proof:* See Appendix B. □

The above proposition can intuitively be interpreted as saying that a sufficient relation is one where different states with the same label are not only indistinguishable on their own, but are actually indistinguishable even by their consequences. Starting from one of two states with same labels, there is no way to ever find out which one of them it was, no matter how much will the agent investigate its environment, compare to the discussion in the end of Section 2.8.

Proposition 4.5 below is an important proposition on which the idea of derived I-spaces and combinatorial filters builds upon (LaValle, 2006, 2012; O'Kane and Shell, 2017), although as far as the authors are aware, in the literature, only the "if"-direction is mentioned. We say that a transition system $(X, U, T)$ is *full*, if for all $x_1 \in X$ and all $u \in U$ there exists at least one $x_2 \in X$ with $(x_1, u, x_2)$.

**Proposition 4.5.** *Suppose $\mathcal{X} = (X, U, T)$ is a transition system. Let $h\colon X \to L$ be a labeling. Then $\mathcal{X}/h$ is an automaton if and only if $\mathcal{X}$ is full and $h$ is sufficient.*

*Proof:* See Appendix B. □

The above proposition brings together the ideas of a quotient, automaton and sufficiency. The idea of the quotient is that indistinguishable states can be in some circumstances considered the same and the idea of an automaton is that it is deterministic. The above proposition says that as far as the agent is concerned, if it equalizes indistinguishable states, then the world looks deterministic from the agent's perspective if and only if the underlying labeling satsifies Definition 4.2.

The sufficiency of an information mapping was introduced in LaValle (2006, Ch 11), and is encompassed by a sufficient labeling in this paper. In the prior context, it has meant that the current sensory perception together with the next action determine the next sensory perception. The elegance with respect to our principle (EA2) is that sufficiency is *not* saying that the agent's internal state corresponds to the environment's state (as is in representational models). Nor is it saying that the agent *predicts* the next action. It is saying, rather, that the agent's current sensation together with a choice of a motor command *determine* the agent's next sensation; and this statement is true only as a statement made about the system from outside, not as a statement which would reside "in the agent." The sensation may carry no meaning at all "about" what is actually "out there." However, if the agent has found a way to be coupled to the environment in a sufficient way, then sensations *begin* to be *about* future sensation. In this way meaning emerges from sensorimotor patterns. This relates to (EA3) and somewhat touches on the topic of perception (EA5). Furthermore, the property of determining future outcomes is related to (EA4) because that is what *skill* is. There is no potential to *reliably* engage with the environment in complex sensorimotor interactions, if the sensations do not *reliably* follow certain historical patterns.

Thus, the notion of sufficiency is considered by us to be of fundamental importance for enactivist-inspired mathematical modeling of cognition. The violation of sufficiency means that the current sensation-action pair does not correlate with the future sensation, making it harder to be attuned to the environment. Having a different sensation following the same pattern can be seen as a primitive notion of a "surprise." This can be seen as aligning with the predictive coding and the free energy principle from neuroscience (Rao and Ballard, 1999; Friston and Kiebel, 2009; Friston, 2010), although our framework leaves the space to a clean non-representational interpretation while this is not obvious for these other frameworks. Does the notion of sufficient labelings capture the same ideas in a more general way? This is an open question for further research.

A generalization of sufficiency is $n$-sufficiency, in which the data of $n$ previous steps is needed to determine the next label. Here, we define an $n$-chain.

**Definition 4.6.** An *$n$-chain* in $\mathcal{X} = (X, U, T)$ is a sequence

$$c = (x_0, u_0, \cdots, x_{n-1}, u_{n-1}, x_n) \in (X \times U)^n \times X$$

such that $x_i \xrightarrow{u} x_{i+1}$ for all $i < n$. If $n = 0$, then by convention $c = (x_n)$. Let $E \subseteq X \times X$ be an equivalence relation. Let $k < n$. We say that two $n$-chains $c = (x_0, u_0, \ldots, x_{n-1}, u_{n-1}, x_n)$, $c' = (x'_0, u'_0, \ldots, x'_{n-1}, u'_{n-1}, x'_n)$ are $(T, E, k)$-equivalent if for all $i < k$, we have $u_i = u'_i$ and $(x_i, x'_i) \in E$. An $\infty$-chain is defined in the same way as $n$-chain, except the sequences are infinite, without the "last" $x_n$.

**Definition 4.7.** For a transition system $\mathcal{X} = (X, U, T)$, an equivalence relation $E$ on $X$ is called *$n$-sufficient* if there are no two $(T, E, n)$-equivalent $n$-chains

$$c = (x_0, u_0, \ldots, x_{n-1}, u_{n-1}, x_n) \text{ and}$$
$$c' = (x'_0, u'_0, \ldots, x'_{n-1}, u'_{n-1}, x'_n)$$

such that $(x_n, x'_n) \notin E$. A labeling $h\colon X \to L$ is called *$n$-sufficient* if $E^h$ is $n$-sufficient (Recall Definition 2.34).

**Proposition 4.8.** *An equivalence relation $E$ is 0-sufficient if and only if there is only one $E$-equivalence class, and a labeling function $h$ is 0-sufficient if and only if it is constant.*

*Proof:* See Appendix B □

**Proposition 4.9.** *An equivalence relation $E$ (resp. a labeling $h$) is sufficient if and only if it is 1-sufficient.*

*Proof:* See Appendix B □

**Proposition 4.10.** *Suppose $n < m$ are natural numbers. If a labeling $h$ is $n$-sufficient, then it is $m$-sufficient. The same holds for equivalence relations.*

*Proof:* See Appendix B □

This enables us to define the degree of insufficiency:

**Definition 4.11.** The *degree of insufficiency* of the labeled automaton $\mathcal{X} = (X, U, \tau, h, L)$ is defined to be the smallest $n$ such that $h$ is $n$-sufficient, if such $n$ exists, and $\infty$ otherwise. Denote the degree of insufficiency of $\mathcal{X}$ by degins($\mathcal{X}$), or degins($h$) if only the labeling needs to be specified and $\mathcal{X}$ is clear from the context.

The intuition is that the larger the degree of insufficiency of an environment $\mathcal{X}$, the harder it is for an agent to be attuned to it. We talk more about the connection between attunement and sufficiency in the following sections.

## 4.2. Minimal sufficient refinements

In this section we prove that the minimal sufficient refinements are always unique (Theorem 4.19). This will follow from a deeper result that the sufficient equivalence relations form a complete sublattice of the lattice of all equivalence relations. This does not hold for $n$-sufficient equivalence relations for $n > 1$ (Example 4.20). We will then explore how the minimal sufficient refinements can be thought of as an enactive perceptual construct that emerges from the body-environment, brain-body, and brain-body-environment dynamics. The idea is that a minimal sufficient refiniment corresponds to an optimal attunement of the agent to the base labeling which corresponds to some minimal information that the agent is interested in the environment, such as death or life, danger or safety information. It is "optimal" by minimality and "attunement" by sufficiency. Our Theorem 4.19 states that such attunement is mathematically unique.

**Definition 4.12.** An equivalence relation $E$ is a *refinement* of equivalence relation $E'$, if $E \subseteq E'$, also denoted $E' \leqslant_r E$. A labeling function $h$ is a refinement of a labeling function $h'$, if $E^h$ is a refinement of $E^{h'}$.

An important interpretation of the concept of a refinement is that a better sensor provides the agent with more information about the environment[5]. Each sensor mapping $h$ induces a partition of $X$ *via* its preimages, and refinement applies in the usual set-theoretic sense to the partitions when comparing sensors mappings. If a sensor mapping $h$ is a refinement of $h'$, then it enables the agent to react in a more refined way to nuances in the environment. Using the partial ordering given by refinements, we obtain the *sensor lattice* (LaValle, 2019).

By a referee's request, let us give a couple of biological examples.

**Example 4.13** (First biological example)**.** There is an accepted theory that primates see red color wavelength, because it enables them to distinguish ripe fruit from non-ripe. Assuming this theory is true, it is an example of a refinement which is to some extent "minimal" and to some extent "sufficient" (of course strictly speaking it is neither – in the same way as there is no ideal circle in the physical world). The minimality is seen in this example, because we perceive other things as red, even if it is completely unnecessary (certainly unnecessary to tell the ripeness of fruits). So we are not distinguishing "too much." On the other hand, perceiving red color is a refinement of ripe/non-ripe which is only detected through stomach ache after the fruit has been already consumed. And it is sufficient in the sense that it is predictive of the original "base" labeling (ripe/non-ripe).

**Example 4.14** (Second biological example)**.** Where our eyes look depends on the position of our head as well as the position of our eyes. Despite this, "looking up" (or "left," "right" etc..) are not ambiguous, even though these can be achieved with virtually infinitely many different head-eye configurations. One way to understand how this invariance could emerge is through minimal sufficient refinements. Suppose at birth, every head-eye configuration is considered as a separate state, but we label them by what we see in any given (stable) situation. A minimal sufficient refinement of that labeling will never distinguish between different states in which the eyes are pointing in the same direction. So then, by learning the minimal sufficient refinements, the agent may learn eye-direction invariance.

## 4.3. Lattice of sufficient equivalence relations

Please refer to Appendix A in the Supplementary material for notations and definitions used in this section.

We will prove in this section that if $(X, U, \tau)$ is an automaton, the sufficient equivalence relations form a complete sublattice of $(\mathcal{E}(X), \subseteq)$. Given an automaton $\mathcal{X} = (X, U, \tau)$, denote by $\mathcal{E}_{\text{suf}}^{U,\tau}(X) \subseteq \mathcal{E}(X)$ the set of sufficient equivalence relations on $X$. When $U$ and $\tau$ are clear from the context, we write just $\mathcal{E}_{\text{suf}}(X) = \mathcal{E}_{\text{suf}}^{U,\tau}(X)$.

**Theorem 4.15.** *Suppose $(X, U, \tau)$ is an automaton and suppose that $\mathcal{E} \subseteq \mathcal{E}_{\text{suf}}(X)$ is a set of sufficient equivalence relations. Then $\bigwedge \mathcal{E}$ and $\bigvee \mathcal{E}$ are sufficient. Thus, $(\mathcal{E}_{\text{suf}}(X), \subseteq)$ is a complete sublattice of $(\mathcal{E}(X), \subseteq)$.*

*Proof:* See Appendix B. □

5 Here we are not talking about contentful or semantic information, but merely about correlational information in the philosophical sense.

Suppose that a labeling $h$ is very important for an agent. For example, $h$ could be "death or life," or it could be relevant for a robot's task. Suppose that $h$ is not sufficient. The robot may want to find a sufficient refinement of $h$. Clearly a one-to-one $h'$ would do. However, assume that the agent has to use resources for distinguishing between states; thus, the fewer distinctions the better. This motivates the following definition. Recall Definition 4.12 of refinements.

**Definition 4.16.** Let $(X, U, T)$ be a transition system and $E_0 \subseteq X \times X$ an equivalence relation. *A minimal sufficient refinement* of $E_0$ is a sufficient equivalence relation $E$ which is a refinement of $E$ such that there is no sufficient $E'$ with $E_0 \leqslant_r E' <_r E$.

Given a labeling $h_0$ of a transition system $(X, U, T)$, a *minimal sufficient refinement* of $h_0$ is a labeling $h$ such that $E^h$ is a minimal sufficient refinement of $E^{h_0}$ (recall Definition 2.34).

**Example 4.17.** Let $\mathcal{X} = (X, U, \tau)$ be an automaton where $X = \{0,1\}^*$, $U = \{0,1\}$ and $\tau(x, b) = x^\frown b$ (concatenation of the binary string $x$ with the bit $b$). Let $h(x) = 1$ if and only if the number of ones and the number of zeros in $x$ are both prime; otherwise $h(x) = 0$. Then the only sufficient refinements of $h$ are one-to-one.

**Example 4.18.** Let $\mathcal{X}$ be as above and let $h\colon X \to \{0,1\}$ be such that if $|x|$ is divisible by 3, then $h(x) = 1$; otherwise, $h(x) = 0$. Then $h$ is not sufficient. Let $h'\colon x \mapsto \{0,1,2\}$ be such that

$$h'(x) \equiv |x| \mod 3.$$

Then $h'$ is a minimal sufficient refinement of $\sigma$.

**Theorem 4.19.** *Consider an automaton $\mathcal{X} = (X, U, \tau)$ and let $E_0$ be an equivalence relation on $X$. Then a minimal sufficient refinement of $E_0$ exists and is unique.*

*Proof:* See Appendix B □

Theorem 4.19 fails, if "automaton" is replaced by "transition system," or if "sufficient" is replaced by "$n$-sufficient" for $n > 1$ (recall Definition 4.7)

**Example 4.20** (Failure of uniqueness for $n$-sufficiency)**.** Let $X = \{0,1,2,3,4,5\}$, $U = \{u_0\}$ and

$$\tau(0,u_0) = 1, \tau(1,u_0) = 2, \tau(2,u_0) = 2,$$

and

$$\tau(3,u_0) = 4, \tau(4,u_0) = 5, \tau(5,u_0) = 5.$$

Let $E_0$ be an equivalence relation on $X$ such that the equivalence classes are $\{0,1,3,4\}$, $\{2\}$ and $\{5\}$. Then this relation is not 2-sufficient, because $(0, u_0, 1, u_0, 2)$ and $(3, u_0, 4, u_0, 5)$ are $(T, E_0, 2)$-equivalent, but 2 and 5 are not $E_0$-equivalent. Let $E_1, E_2 \subseteq E_0$ be equivalence relations with equivalence classes as follows:

$$E_1 :\{0,1\},\{3,4\},\{2\},\{5\},$$

$$E_2 :\{0,4\},\{1,3\},\{2\},\{5\}.$$

Then $E_1$ and $E_2$ are refinements of $E_0$. They are both 2-sufficient, because there doesn't exist any $(T, E_1, 1)$ or $(T, E_2, 1)$ equivalent 2-chains. They are also both $\leqslant_r$-minimal with this property which can be seen from the fact that they are actually $\leqslant_r$-minimal refinements of $E_0$ as equivalence relations (not only as sufficient ones).

**Example 4.21** (Failure of uniqueness for transition systems)**.** Let $X = \{0,1,2,3,4\}$, $U = \{u_0\}$ and $T = \{(0, u_0, 3), (2, u_0), 4\}$. Let $E_0$ be the equivalence relation with the equivalence classes $\{0,1,2\}$, $\{3\}$ and $\{4\}$. Then $E_0$ is not sufficient, because $(0,2) \in E_0$, but $(3,4) \notin E_0$. Let $E_1$ and $E_2$ be the refinements of $E_0$ with the following equivalence classes:

$$E_1 :\{0,1\},\{2\},\{3\},\{4\},$$

$$E_2 :\{0\},\{1,2\},\{3\},\{4\}.$$

Now it is easy to see that both $E_1$ and $E_2$ are sufficient refinements of $E_0$, and by a similar argument as in Example 4.20 they are both minimal. The reason why this is possible is the odd behavior of the state 2 which doesn't have out-going connections. Such odd states are the reason why the decision problem "Does there exist a sufficient refinement with $k$ equivalence classes?" is NP-complete for finite transition systems (O'Kane and Shell, 2017).

*Remark.* It is worth noting that Theorems 4.15 and 4.19 do not assume anything about the cardinality of $X$ or of $U$, other structure on them (such as metric or topology) nor anything about the function $\tau$ or the relation $E_0$. Keeping in mind potential applications in robotics, $X$ and $U$ could be, for instance, topological manifolds, and $\tau$ a continuous function, or $X$ could be a closed subset of $\mathbb{R}^n$, $U$ discrete and $\tau$ a measurable function, or any other combination of those. In each of those cases, the sublattice of sufficient equivalence relations is complete, as per Theorem 4.15, and every equivalence relation $E_0$ on $X$ admits a unique minimal sufficient refinement as per Theorem 4.19.

Recall Definition 2.10 of an equivalence relation preserving function. We say that an equivalence relation $E$ on $X$ is *closed under* $f\colon X \to X$ if for all $x \in X$, we have $(x, f(x)) \in E$. If $E$ is closed under $f$, then $f$ is $E$-preserving: given $(x, x') \in E$, we have $(x, f(x)), (x', f(x')) \in E$, because $E$ is closed under $f$. Now by transitivity of $E$ we have $(f(x), f(x')) \in E$, so $f$ is $E$-preserving.

**Definition 4.22.** Let $f\colon X \to X$ be a bijection. The induced *orbit equivalence relation* is the relation $E_f$ on $X$ defined by $(x, x') \in E_f \iff (\exists n \in \mathbb{Z})(f^n(x) = x')$, in which $f^n(x)$ is defined by induction as: $f^0(x) = x$, $f^{n+1}(x) = f(f^n(x))$, $f^{n-1}(x) = f^{-1}(f^n(x))$.

**Theorem 4.23.** *If $f$ is an automorphism of the automaton $(X, U, \tau)$, then $E_f$ is a sufficient equivalence relation.*

*Proof:* See Appendix B □

**Theorem 4.24.** *Let $\mathcal{X} = (X, U, \tau)$ be an automaton and $E$ be an equivalence relation on $X$. Suppose $f: X \to X$ is an automorphism such that $E$ is closed under $f$. Let $E'$ is the minimal sufficient refinement of $E$. Then $E'$ is closed under $f$ and $E \leqslant_r E' \leqslant_r E_f$.*

*Proof:* See Appendix B □

**Example 4.25.** Consider the environment which is a one-dimensional lattice of length five, $E = \{-2, -1, 0, 1, 2\}$, in which the corners "smell bad"; thus, we have a sensor mapping $h: E \to S$, $S = \{0, 1\}$ defined by $h(n) = 0 \iff |n| = 2$; see Figure 6A. Consider two agents in this environment. Both are equipped with the same $h$ sensor, but their action repertoires differ. Both have two possible actions. One has actions $L$ = "move left one space" and $R$ = "move right one space," and the other one has actions $T$ = "turn 180 degrees" and $F$ = "go forward one space." Let $M_0 = \{L, R\}$ and $M_1 = \{T, F\}$. Thus, these agents have a slight difference in embodiment. Although both of them can move to every square of the lattice in a very similar way (almost indistinguishable from the outside perspective), we will see that the differences in embodiment will be reflected in that the minimal sufficient refinements will produce non-equivalent "categorizations" of the environment. The structures that emerge from these two embodiments will be different. These agents *enact* different environments, although physically the environments are the same, as congruent with tenet (EA3).

First, we define the SM-systems that model these agents' embodiments in $E$. The first agent does not have orientation. It can be in one of the five states, and the state space is $X_0 = E$ (Figure 6B). For the second agent, the effect of the $F$ action depends on the orientation of the agent (pointing left or pointing right). Thus, there are ten different states the agent can be in, yielding $X_1 = E \times \{-1, 1\}$ (Figure 6C). The effects of motor outputs are specified completely ($L$ means moving left, and so on), whereas the agent's internal mechanisms are left completely open, so our systems will be quasifilters. According to Remark 2.18, we can work with a labeled automaton instead. Hence, let $\tau_0: X_0 \times M_0 \to X_0$ be defined by $\tau_0(x, L) = \max(x - 1, -2)$ and $\tau_0(x, R) = \min(x + 1, 2)$. For the other agent, let $\tau_1((x, b), T) = (x, -b)$ and $\tau_1((x, b), F) = (\min(\max(x + b, -2), 2), b)$. Now we have labeled automata $\mathcal{X}_0 = (X_0, M_0, \tau_0, h, S)$ and $\mathcal{X}_1 = (X_1, M_1, \tau_1, h, S)$.

It is not hard to see that the one-to-one map $h_0: X_0 \to \{-2, -1, 0, 1, 2\}$ with $h_0(x) = x$ is a sufficient refinement of $h$ which is minimal (see Figure 7A). Thus, every state needs to be distinguished by the agent for it to be possible to determine the following sensation from the current one. The derived information space automaton $\mathcal{X}_0/h_0$ isomorphic to $\mathcal{X}_0$ (Proposition 2.36).

For the second automaton, consider the labeling $h_1: X_1 \to \{-2, -1, 0, 1, 2\}$ defined by $h_1(x, b) = b \cdot x$ (see Figure 7C).

**Claim.** $h_1$ is a minimal sufficient refinement of $h$ in $\mathcal{X}_1$.

*Proof:* See Appendix B □

Both minimal sufficient labelings, $h_0$ and $h_1$ have five values; thus, they categorize the environment into five distinct state-types. However, the resulting derived information spaces are different in the sense that the quotients $\mathcal{X}_0/h_0$ and $\mathcal{X}_1/h_1$ are not isomorphic; compare Figure 7B with Figure 7D.

**Example 4.26.** Figure 8A shows a filtering example from Tovar et al. (2014). More complex versions have been studied more recently in O'Kane and Shell (2017), and are found through automaton minimization algorithms and some extensions. It can be shown that this example's four-state derived information space depicted on Figure 8B corresponds to the unique minimal sufficient refinement of the labeling that only distinguishes between "are in the same region" and "are not in the same region." To see this, first note that this labeling is sufficient (since it can be represented as an automaton, this follows from Theorem 4.5). It follows from Theorem 4.19 that if this labeling is not minimal, then there is a minimal one which is strictly coarser, and so can be obtained by merging the states in the automaton of Figure 8B. This is impossible: the state $T$ cannot be merged with anything because it violates the base-labeling; if, say $D_a$ and $D_c$, are merged, then transition $a$ will lead to inconsistency as it can lead either to $D_b$ (from $D_c$) or to $T$ (from $D_a$). This proves that this derived information space is indeed minimal sufficient, and by Corollary 4.19 there are no others up to isomorphism.

## 4.4. Computing sufficient refinements

This section sketches some computational problems and presents computed examples. The problem of computing the minimal sufficient refinement in some cases reduces to classical deterministic finite automaton (DFA) minimization, and in other cases it becomes NP-hard (O'Kane and Shell, 2017).

Consider an automaton $(X, M, \tau)$ and a labeling function $h_0$, and the corresponding labeled automaton described using the quintuple $(X, M, \tau, h_0, L)$. Suppose that the automaton $(X, M, \tau)$ corresponds to that of an body-environment system. Hence, $X$ corresponds to the states of this coupled system. Suppose $h_0$ is not sufficient and consider the problem of computing a (minimal) sufficient refinement of $h_0$, that is, the coarsest refinement of $h_0$ that is sufficient.

Despite the uniqueness of the minimal sufficient refinement of $h_0$ (by Corollary 4.19), we can argue that the formulation of the problem, in particular, the input, can differ based on the level at which we are addressing the problem (for example, global perspective, agent perspective or something in between). Since the labeled automaton corresponding to an agent-environment coupling is described from a global perspective, the input to an

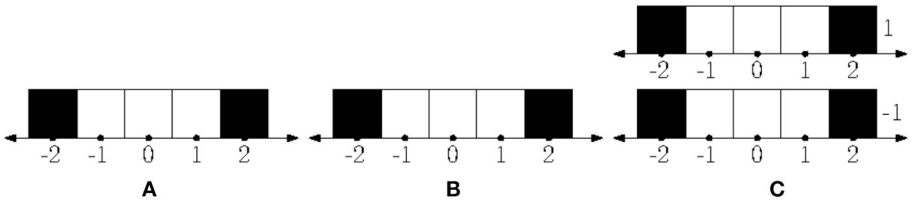



FIGURE 6
**(A)** One-dimensional lattice environment described in Example 4.25. **(B)** State space of the agent 0. **(C)** State space of the agent 1. The states for which the value of the sensor mapping is 0 are shown in black.

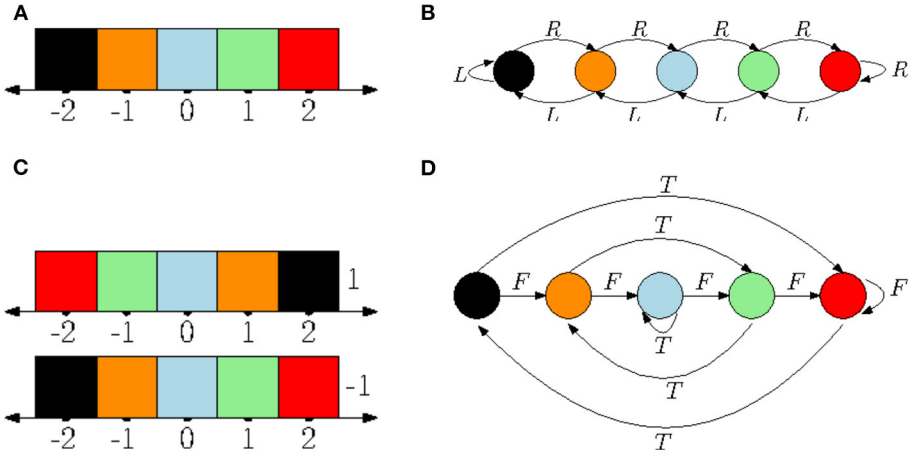



FIGURE 7
**(A)** State space of the agent 0 categorized by $h_0$, states that belong to the same class are colored with the same color. **(B)** Resulting quotient $\mathcal{X}_0/h_0$ for the agent 0. **(C)** State space of the agent 1 categorized by $h_1$, states that belong to the same class are colored with the same color. **(D)** Resulting quotient $\mathcal{X}_1/h_1$ for the agent 1.

algorithm that addresses the problem from this perspective is the labeled automaton $\mathcal{A} = (X, \tau, M, h_0, L)$ itself. Then, the problem is defined as given $\mathcal{A}$ compute $\mathcal{A}' = (X, M, \tau, h, L)$ such that $h$ is the minimal sufficient refinement of $h_0$.

A special case of this problem from the global perspective occurs if the preimages of $h_0$ partition $X$ in two classes which can be interpreted as the "accept" and "reject" states, for example, goal states at which the agent accomplishes a task and others. Furthermore, suppose that the initial state of the agent is known to be some $x_0 \in X$. Then, computing a minimal sufficient refinement becomes identical to minimization of a finite automaton, that is, given a DFA $(X, M, \tau, x_0, F)$ in which $x_0$ is the initial state and $F$ is the set of accept states find $(X', M, \tau', x_0', F')$ such that no DFA with fewer states recognizes the same language. Existing algorithms, for example Hopcroft (1971), can be used to compute a minimal automaton.

Here, we also consider this problem from the agent's perspective for which the information about the environment states is obtained through its sensors, more generally, through a labeling function. Note that by agent's perspective we do not necessarily imply that the agent is the one making the computation (or any computation) but it means that no further information can be gathered regarding the environment other than the actions taken and what is sensed by the agent. At this level we address the following problem; given a set $M$ of actions, a domain $X$, and a labeling function $h_0$ defined on $X$, compute the minimal sufficient refinement of $h_0$. The crux of the problem is that unlike the global perspective described above, the labeled automaton $\mathcal{A}$ is not given, in particular, the state transitions are not known a priory. Instead, the information regarding the state transitions can only be obtained locally by means of applying actions and observing the outcomes, that is, through sensorimotor interactions. Hence, the current body-environment state is also not observable. To show that an algorithm exists to compute a sufficient refinement of $h_0$ at this level, we propose an iterative algorithm (Algorithm 1) that explores $X$ through agent's actions and sensations by keeping the history information state, that is, the history of actions and sensations (labels). We then show, by empirical results, that the

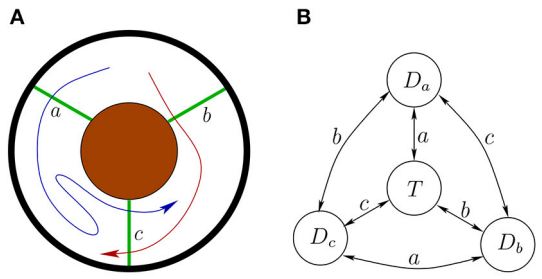



FIGURE 8
**(A)** Two point-sized independent bodies move along continuous paths in an annulus-shaped region in the plane. There are three sensor beams, *a*, *b*, and *c*. When each is crossed by a body, its corresponding symbol is observed. Based on receiving a string of observations, the task is to determine whether the two bodies are together in the same region, with no beam separating them. **(B)** The minimal filter as a transition system has only 4 states: *T* means that they are together, and $D_x$ means that are in different regions but beam *x* separates them. Each transition is triggered by the observation when a body crosses a beam.

sufficient refinement computed by Algorithm 1 is minimal for the selected problem.

```
1: Input: h_0, l_0, M
2: Initialize: H ← ∅, h ← h_0, s ← s_0
3: for each step do
4:    m ← policy(s)
5:    apply action m and obtain resulting s'
6:    add (s, m, s') to H
7:    if ∃(s, m, s'') ∈ H such that s' ≠ s'' then
8:       h ← split(h, s)
9:    if there are labels that can be merged then
10:      h ← merge(h, H, h_0)
11:   s ← s'
```

Algorithm 1.

The functioning of Algorithm 1 is as follows. Starting from an initial sensation $s_0 = h(x_0)$, the agent moves by taking an action[6] given by the mapping policy$: L \rightarrow M$. Particularly, we used a fixed policy which samples an action $m$ from a uniform distribution over $M$ for each $s \in S$. In principle, any policy that ensures all states that are reachable from $x_0$ will be visited infinitely often should be enough. The history information state is implemented as a list, denoted by $H$, of triples $(s, m, s')$ such that $s = h(x)$ and $s' = h(x')$ in which $x' = \tau(x, m)$. At each step, it is checked whether the current sensation is consistent with the history (Line 7). Current sensation is inconsistent with the history if there exists a triple $(s, m, s'')$ in the history such that $s' \neq s''$. If it is not consistent then the label is split, which means that $h^{-1}(s)$ is partitioned into two parts $P$ and $Q$. In particular, we apply a balanced random partitioning, that is, we select $P$ and $Q$ randomly from a uniform distribution over the partitions of $h^{-1}(s)$ that have two elements with balanced cardinalities. The labeling function is updated by a split operation as

$$h(x) := \begin{cases} s_Q & \text{if } x \in Q \\ s_P & \text{if } x \in P \\ h(x), & \text{otherwise.} \end{cases}$$

Recall that labels or subscripts do not carry any meaning from the agent's perspective.

Even a trivial strategy that splits the preimage of the label seen at each step would succeed computing a sufficient refinement. However, this would result in $h$ being a one-to-one mapping. Hence, the finest possible refinement. Splitting only at the instances when an inconsistency is detected might reach a coarser refinement that is sufficient but there might be more equivalence classes than the ones induced by the minimal sufficient refinement of $h_0$. Therefore, a merge operation is introduced (Line 10). Let $s$ and $s'$ be two distinct labels for which $\exists s'' \in h_0[X]$ such that $h^{-1}(s) \subseteq h_0^{-1}(s'')$ and $h^{-1}(s') \subseteq h_0^{-1}(s'')$. Let $t$ denote a triple in $H$ and let $t_k$, $k = 1, 2, 3$, denote the $k^{th}$ element of that triple. Suppose $s' = s$, if there are at least $N$ number of triples in $H$ such that for each triple $t$, $(t_1, t_2) = (s, m)$ and $\forall m \in M$ and $\forall t, t' \in H$ such that $(t_1, t_2) = (t'_1, t'_2) = (s, m)$ it is true that $t_3 = t'_3$ then labels $s$ and $s'$ are merged. The merge procedure goes through all labels and updates $h$ as

$$h(x) := \begin{cases} s & \text{if } h(x) \in \{s, s'\} \\ h(x) & \text{otherwise.} \end{cases}$$

for each pair of labels $s$ and $s'$ that satisfies the aforementioned condition. Note that in principle, one can merge two labels regardless of the number of occurrences in the history. However, we noticed that this can result in oscillatory behaviour between split and merge operations especially for states that are reached less frequently. At present, we considered $N$ as a tunable parameter and we know that it depends on the cardinality of the state space $X$ such that larger the number of states, larger $N$ should be. The problem of defining $N$ as a function of the problem description remains open.

In the following, we present an illustrative example to show the practical implications of the previously introduced concepts in Section 4.2. In particular, we show how a simple algorithm like Algorithm 1 can be used by a computing unit which relies only on the sensorimotor interactions of an agent to further categorize the environment such that there are no inconsistencies in terms of the actions taken by the agent and the resulting sensations with respect to an initial categorization induced by $h_0$ (Figure 9C).

6 This can either be in a real environment or in a realistic simulation.

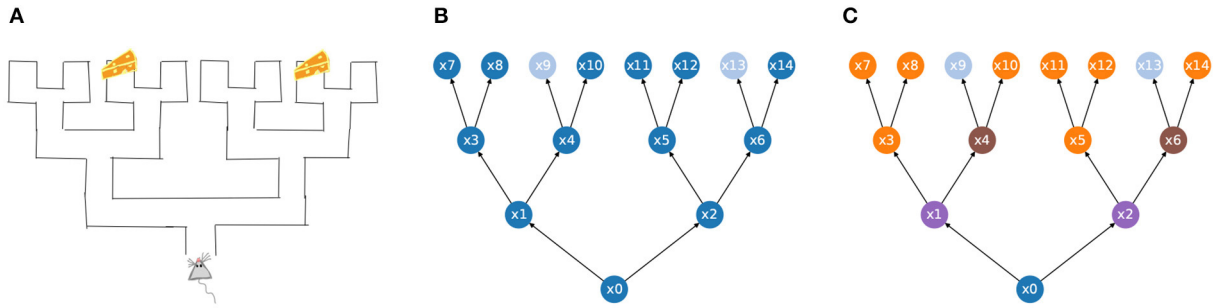



FIGURE 9
**(A)** Cheese maze defined in Example 4.27 **(B)** Labeled automaton with initial labeling $h_0$ corresponding to the cheese-maze example. **(C)** Minimal sufficient refinement of $h_0$. Self-loops at the leaf nodes are not shown in the figure.

**Example 4.27.** Consider an agent (a mouse) that is placed in a maze where certain paths lead to cheese and others do not (see Figure 9A). At each intersection the agent can go either left or right and it can not go back. Hence, at each step the agent takes one of the two actions; go right or go left. Figure 9B shows the corresponding automaton with 15 states describing the agent-environment system together with the initial labeling $h_0$ that partitions the state space into states in which the agent has reached a cheese (light blue) and others (dark blue). The initial state $x_0$ is when the agent is at the entrance of the maze. Once the end of the maze is reached (a leaf node) the state does not change regardless of which action is taken. After a predetermined number of steps the system reverts back to the initial state, similar to an episode in the reinforcement learning terminology (see, for example, Sutton and Barto, 2018). However, despite the system going back to the initial state the history information state still includes the prior actions and sensations. Figure 10 reports the updates of $h$, initialized at $h_0$, by Algorithm 1 being run for 1,000 steps. It converged to a final labeling $h$ (Figure 10R), that is the minimal sufficient refinement of $h_0$, in 435 steps. For 20 initializations of Algorithm 1 for the same problem, on average, it took 364 steps to converge to a minimal sufficient refinement of $h_0$ (Figure 9C).

We have also applied the same algorithm to variations of this example with different depths of maze and different number of cheese and cheese placements (varying $h_0$). Empirical evidence shows that the same algorithm was capable of consistently finding the minimal sufficient refinement of the initial labeling. However, it is likely that it might fail for more complicated problems, for example, when the number of actions are significantly larger. It remains an open problem finding a provably correct algorithm for computing the minimal sufficient refinement of $h_0$ from the agent's perspective.

## 4.5. Sufficiency for coupled SM-systems

Section 2 introduced SM-systems, including the special class of quasifilters. We showed that quasifilters can be thought of as labeled transition systems, and we worked with such systems in Sections 4, 4.4. Let us see how do the concepts introduced in those sections work for SM-systems. We also defined *coupling* of SM-systems (Definition 2.22), but we have not defined what it means for a coupling to be "good." We will use sufficiency to approach this subject.

Let $\mathcal{E} = (E, (S \times M), T)$ and $\mathcal{I} = (I, S \times M, B)$ be SM-systems. We think intuitively of $\mathcal{E}$ as "the environment" and $\mathcal{I}$ as the "agent," even though they share the set of sensorimotor parameters $S \times M$. When is the coupling $\mathcal{E} * \mathcal{I}$ "successful"? Given another $\mathcal{I}' = (I', S \times M, B')$, how can we compare $\mathcal{I}$ and $\mathcal{I}'$ in the context of $\mathcal{E}$? The coupled system $\mathcal{E} * \mathcal{I}$ is not labeled; therefore, we cannot apply the definition of sufficiency. However, as soon as we apply some labeling to it, we can. There are many different ways to do it, intuitively corresponding to the "agent's perspective," the "environment's perspective" and a "god's perspective" (or "global perpsective").

The first one is the labeling $h\colon E \times I \to I$, which is the projection to the right coordinate, $h_I(e, i) = i$. The second one is the projection to the left coordinate $h_E(e, i) = i$, and the third one is the labeling of states by themselves, $h_G(e, i) = (e, i)$. Clearly, $h_G$ is a refinement of both $h_E$ and $h_I$. Yet another option is to use the sensory data as labelings, which is a coarser labeling than $h_I$. Or perhaps there was already a labeling $h\colon E \to S$ to begin with, so then we can ask about the property of $\hat{h}\colon E \times I \to S$ defined by $\hat{h}(e, i) = h(e)$. We focus on what we called the agent's perspective, $h_I$, for the rest of this section.

Recall Definition 4.11 of the degree of insufficiency. Given SM-systems $\mathcal{E}$ (environment) and $\mathcal{I}$ (agent), we can ask what is the degree of insufficiency of $h_I$ in $\mathcal{E} * \mathcal{I}$? The smaller the

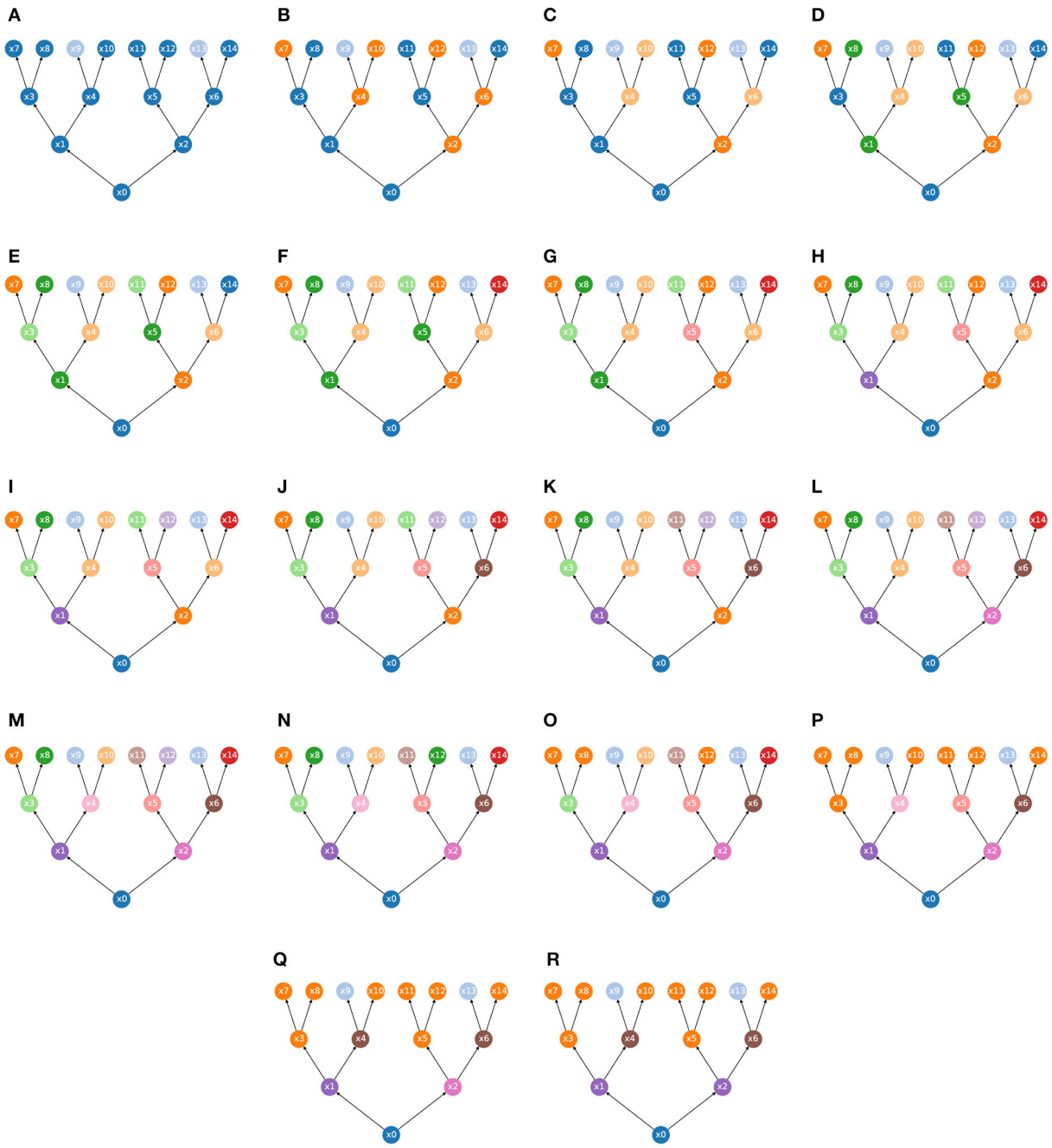



FIGURE 10
**(A)** Labeled automaton with labefigure/ling function $h = h_0$; same colored states belong to the figure/same equivalence class. **(B–Q)** Updating $h$ by Algorithm 1 through splitting and merging of the labels. **(R)** Labeled automaton with the labeling function $h$ that is the minimal sufficient refinement of $h_0$.

degree, the better the agent is attuned to the environment. This says something about the way in which the agent is adapted or attuned to the environment without attributing contentful states or representations to the agent in alignment with (EA2) and (EA4).

Let $\mathcal{E}$, $\mathcal{I}$, and $\mathcal{I}'$ be SM-systems. When is degins($\mathcal{E} * \mathcal{I}, h_I$) < degins($\mathcal{E} * \mathcal{I}', h_{I'}$)? Of course, if $\mathcal{I}$ is fully constrained (Definition 2.31), then degins($\mathcal{E} * \mathcal{I}$) = $\infty$. This corresponds to the agent never engaging in any sensorimotor interaction with the environment. No wonder that it can always "predict" the

result of such passive existence. Assume, however, that there some constraints on the coupling. For example, we may demand that the agent must regularly visit states of some particular type to survive. Subject to such constrains, what can we say about degins($\mathcal{E} * \mathcal{I}$)? This seems to be a good preliminary notion[7] of attunement.

## 5. Discussion

In the introduction we defined our basic enactivist tenets:

(EA1) Embodiment and the inseparability of the brain-body-environment system,
(EA2) Grounding in sensorimotor interaction patterns, not in contentful representations.
(EA3) Emergence from embodiment, enactment of the world,
(EA4) Attunement, adaptation, and skill as possibilities to reliably engage in complicated patterns of activity with the environment.
(EA5) Perception as sensorimotor skills.

We developed a model of sensorimotor systems and coupling for which the purpose is to account for cognition mathematically, but in congruence with the principles (EA1)–(EA5). The principle (EA1) is intrinsic in the ways SM-systems are supposed to model brain-body and body-environment dynamics. The central ingredient is the control set $S \times M$ in all of those systems which include "motor" and "sensory" part; it is *impossible* in our framework to model say the environment without acknowledging the way in which the body is *part of* it. The approach that the actions of an agent depend solely on the history of its sensorimotor interactions with the environment, our approach is well in the scope of (EA2). We do not assume any representational or symbolic content possessed by the SM-systems. We do not evaluate them normatively by the "correctness" of their internal states, but rather by the ways in which they are, or can be, coupled to the environment and whether their sensory apparatus generates a sufficient sensor mapping or not. Coupling of SM-systems is defined so that two systems constrain each other. Thus, when an agent is coupled to the environment, they constrain each other, thereby creating new global properties of the body-environment system.

The principle (EA4) is mostly discussed in connection with minimal sufficient refinements. Given a labeling, or a categorization, or an equivalence relation on the state space, one can ask how well does this labeling "predict itself." The interpretation of this labeling can be anything from a sensor mapping to the labeling of environmental states by the internal states of the agent which coincide with them (this is not representation, this is mere co-occurence; see enactivist interpretation of the place cells in Hutto and Myin (2017) for comparison). A sufficient sensor mapping can be achieved in many different ways. In Section 4.4 we present a way in which the agent "develops" new sensors to be better attuned to the environment and in that way finds a sufficient sensor mapping. Another way for the agent would be to learn to act in a way that excludes "unpredictability." Both are examples of situations where the agent "structures" its own body-environment reality and gains skill. Finally, perception (EA5) can be understood as sensorimotor patterns on a microlevel. On the other hand, the agent engage in a sensorimotor activity locally without making big moves, such as moving the eyes without moving the body. The result of such sensorimotor interaction is another labeling function on a macro level.

In this paper, we not only presented mathematical definitions, but proved a number of propositions and theorems about them. There would be (and we hope there will be!) much more of them, but they did not fit in this expository work for which the main purpose was to demonstrate the connection of the mathematics in question with the enactive philosophy of mind.

We have already developed more concepts and theorems on top of this framework, including notions of *degree of insufficiency*, *universal covers*, *hierarchies*, and *strategic sufficiency*, but these are omitted here due to space limitations.

In other, more mathematical work, we plan to concentrate on working out mathematical and logical details of the proposed theory as well as applying the ideas to fundamental questions in robotics and autonomous systems, control theory, machine learning, and artificial intelligence.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

## Author contributions

VW, BS, and SL developed the mathematical theory together over the past 2 years after extensive collaborative sessions. The primary author is VW, who wrote the most among the authors. VW contributed more to mathematical proofs. BS contributed more to computation. In addition to individual contributions, SL also played a supervisory role. All authors contributed to writing.

## Funding

This work was supported by a European Research Council Advanced Grant (ERC AdG, ILLUSIVE: Foundations of Perception Engineering, 101020977), Academy of Finland (projects PERCEPT 322637, CHiMP 342556), and Business

7 Further research will indicate how much of this will be accepted by the most radical enactivists.

Finland (project HUMOR 3656/31/2019). All authors are with the Center for Ubiquitous Computing, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland.

## Acknowledgments

The author VW wishes to thank Dr. Otto Lappi for numerous discussion concerning enactivist and other agency which helped in shaping many of the ideas of this paper.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2022.846982/full#supplementary-material

## References

Başar, T., and Olsder, G. J. (1995). *Dynamic Noncooperative Game Theory, 2nd Edn*. London: Academic.

Bertsekas, D. P. (2001). *Dynamic Programming and Optimal Control, Vol. I, 2nd Edn*. Belmont, MA: Athena Scientific.

Blum, M., and Kozen, D. (1978). "On the power of the compass (or, why mazes are easier to search than graphs)," in *Proceedings Annual Symposium on Foundations of Computer Science* (Ann Arbor, MI), 132–142.

Choset, H., Lynch, K. M., Hutchinson, S., Kantor, G., Burgard, W., Kavraki, L. E., et al. (2005). *Principles of Robot Motion: Theory, Algorithms, and Implementations*. Cambridge, MA: MIT Press.

Donald, B. R. (1995). On information invariants in robotics. *Artif. Intell. J.* 72, 217–304. doi: 10.1016/0004-3702(94)00024-U

Donald, B. R., and Jennings, J. (1991). "Sensor interpretation and task-directed planning using perceptual equivalence classes," in *Proceedings 1991 IEEE International Conference on Robotics and Automation* (Sacramento, CA: IEEE), 190–197.

Erdmann, M. A. (1993). Randomization for robot tasks: using dynamic programming in the space of knowledge states. *Algorithmica* 10, 248–291. doi: 10.1007/BF01891842

Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: OUP Oxford.

Fraigniaud, P., Ilcinkas, D., Peer, G., Pelc, A., and Peleg, D. (2005). Graph exploration by a finite automaton. *Theor. Comput. Sci.* 345, 331–344. doi: 10.1016/j.tcs.2005.07.014

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Fuchs, T. (2020). The circularity of the embodied mind. *Front. Psychol.* 11, 1707. doi: 10.3389/fpsyg.2020.01707

Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford: Oxford University Press.

Gallagher, S. (2018). *Decentering the Brain: Embodied Cognition and the Critique of Neurocentrism and Narrow-Minded Philosophy of Mind*. Wollongong, NSW: Constructivist Foundations.

Gallistel, C. R., and King, A. (2009). *Memory and the Computational Brain*. Malden, MA: Wiley-Blackwell.

Ghallab, M., Nau, D., and Traverso, P. (2004). *Automated Planning: Theory and Practice*. San Francisco, CA: Morgan Kaufman.

Goranko, V., and Otto, M. (2007). "5 model theory of modal logic," in *Handbook of Modal Logic, volume 3 of Studies in Logic and Practical Reasoning*, eds P. Blackburn, J. Van Benthem, and F. Wolter (Cambridge: Elsevier), 249–329.

Hager, G. D. (1990). *Task-Directed Sensor Fusion and Planning*. Boston, MA: Kluwer.

Hopcroft, J. (1971). "An n log n algorithm for minimizing states in a finite automaton," in *Theory of Machines and Computations* (Haifa: Elsevier), 189–196.

Hutto, D. D., and Myin, E. (2012). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, MA: MIT Press.

Hutto, D. D., and Myin, E. (2017). *Evolving Enactivism*. Cambridge, MA: MIT Press.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artif. Intell. J.* 101, 99–134. doi: 10.1016/S0004-3702(98)00023-X

Kechris, A. S. (1994). *Classical Descriptive Set Theory, Vol. 156*. Verlag: Springer-Verlag; Graduate Texts in Mathematics.

Kumar, P. R., and Varaiya, P. (1986). *Stochastic Systems*. Englewood Cliffs, NJ: Prentice-Hall.

LaValle, S. M. (2006). *Planning Algorithms*. Cambridge, U.K: Cambridge University Press.

LaValle, S. M. (2012). *Sensing and Filtering: A Fresh Perspective Based on Preimages and Information Spaces, volume 1, 4 of Foundations and Trends in Robotics Series*. Delft: Now Publishers.

LaValle, S. M. (2019). "Sensor lattices: structures for comparing information feedback," in *2019 12th International Workshop on Robot Motion and Control (RoMoCo)* (Poznan: IEEE), 239–246.

Lozano-Pérez, T., Mason, M. T., and Taylor, R. H. (1984). Automatic synthesis of fine-motion strategies for robots. *Int. J. Rob. Res.* 3, 3–24. doi: 10.1177/027836498400300101

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

O'Kane, J. M., and LaValle, S. M. (2008). Comparing the power of robots. *Int. J. Rob. Res.* 27, 5–23. doi: 10.1177/0278364907082096

O'Kane, J. M., and Shell, D. A. (2017). Concise planning and filtering: hardness and algorithms. *IEEE Trans. Autom. Sci. Eng.* 14, 1666–1681. doi: 10.1109/TASE.2017.2701648

O'Regan, J. K., and Block, N. (2012). Discussion of j. kevin o'regan's "why red doesn't sound like a bell: understanding the feel of consciousness". *Rev. Philos. Psychol.* 3, 89–108. doi: 10.1007/s13164-012-0090-7

O'Regan, J. K., and Noëe, A. (2004, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–973. doi: 10.1017/S0140525X01000115

Paolo, E. A. D. (2018). "The enactive conception of life," in *The Oxford Handbook of 4E Cognition*, eds A. Newen, L. de Bruin, and S. Gallagher (Oxford: Oxford University Press), 71–95.

Pezzulo, G., Barsalou, L., Cangelosi, A., Fischer, M., Spivey, M., and McRae, K. (2011). The mechanics of embodiment: a dialog on embodiment and computational modeling. *Front. Psychol.* 2, 5. doi: 10.3389/fpsyg.2011.00005

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind Lang.* 31, 3–36. doi: 10.1111/mila.12093

Roy, N., and Gordon, G. (2003). "Exponential family PCA for belief compression in POMDPs," in *Proceedings Neural Information Processing Systems* Vancouver, BC.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge, U.K: Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1952). "Presentation of a maze-solving machine," in *Transaction of the 8th Cybernetics Conference* (Josiah Macy, Jr., Foundation), 173–180.

Suomalainen, M., Nilles, A. Q., and LaValle, S. M. (2020). "Virtual reality for robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Las Vegas, NV: IEEE), 11458–11465.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. Cambridge, MA: MIT Press.

Tovar, B., Cohen, F., Bobadilla, L., Czarnowski, J., and LaValle, S. M. (2014). Combinatorial filters: sensor beams, obstacles, and possible paths. *ACM Trans. Sens. Networks* 10, 2594767. doi: 10.1145/2594767

Tschacher, W., and Dauwalder, J. P. (2003). *The Dynamical Systems Approach to Cognition: Concepts and Empirical Paradigms Based on Self-organization, Embodiment, and Coordination Dynamics*. Singapore: World Scientific.

Varela, F., Rosch, E., and Thompson, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fcomp.2022.896465

Check for updates

OPEN ACCESS

EDITED BY
Inês Hipólito,
Humboldt University of Berlin, Germany

REVIEWED BY
Christian Spiros Motsenigou Kronsted, University of Memphis, United States
Alejandra Ciria,
Facultad de Psicología, Universidad Nacional Autónoma de México, Mexico

*CORRESPONDENCE
James Garner
james.garner@auckland.ac.nz

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

RECEIVED 15 March 2022
ACCEPTED 26 September 2022
PUBLISHED 17 October 2022

CITATION
Garner J and Egbert MD (2022) Embodiment enables non-predictive ways of coping with self-caused sensory stimuli.
*Front. Comput. Sci.* 4:896465.
doi: 10.3389/fcomp.2022.896465

COPYRIGHT
© 2022 Garner and Egbert. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Embodiment enables non-predictive ways of coping with self-caused sensory stimuli

James Garner[1]* and Matthew D. Egbert[1,2]

[1]Artificial Life and Minds Lab, School of Computer Science, University of Auckland, Auckland, New Zealand, [2]Te Ao Mārama–Centre for Fundamental Inquiry, The University of Auckland, Auckland, New Zealand

Living systems process sensory data to facilitate adaptive behavior. A given sensor can be stimulated as the result of internally driven activity, or by purely external (environmental) sources. It is clear that these inputs are processed differently—have you ever tried tickling yourself? Self-caused stimuli have been shown to be attenuated compared to externally caused stimuli. A classical explanation of this effect is that when the brain sends a signal that would result in motor activity, it uses a copy of that signal to predict the sensory consequences of the resulting motor activity. The predicted sensory input is then subtracted from the actual sensory input, resulting in attenuation of the stimuli. To critically evaluate the utility of this predictive approach for coping with self-caused stimuli, and investigate when non-predictive solutions may be viable, we implement a computational model of a simple embodied system with self-caused sensorimotor dynamics, and use a genetic algorithm to explore the solutions possible in this model. We find that in this simple system the solutions that emerge modify their behavior to shape or avoid self-caused sensory inputs, rather than predicting these self-caused inputs and filtering them out. In some cases, solutions take advantage of the presence of these self-caused inputs. The existence of these non-predictive solutions demonstrates that embodiment provides possibilities for coping with self-caused sensory interference without the need for an internal, predictive model.

KEYWORDS

**sensory attenuation, embodiment, evolutionary robotics, ego-noise, self-other distinction, sensorimotor feedback, computational model, prediction**

## 1. Introduction

The remarkable adaptive behavior displayed by living organisms would not be possible without the capacity to respond to sensory stimuli appropriately. The same sensors can be stimulated due to external (environmental) causes, as well as by internally driven activity. Intuitively, it seems like responding appropriately must require distinguishing the two. We can hear sounds in the world around us, but we can also hear our own voice when talking, and our own footsteps when walking. We can see our environment, but we can also see our own bodies. Not only do we perceive both the world and the results of our own actions, but the exact same sensory stimulus can be caused by an external event, or by our own activity. For example the sight of a hand being waved before your eyes could be your own hand or a friend snapping you out of a daydream.

However, we typically have no trouble telling the difference. Indeed, the phenomenology of a self-caused stimulus can be very different from that of an externally caused one. A great example of this is the sensation of touch, which can reduce you to helpless laughter when externally applied—but trying to tickle yourself just isn't the same! (Blakemore et al., 2000). Understanding exactly how these inputs are processed differently can facilitate building artificial systems as capable and flexible as living ones.

One concrete way this has been studied is in research on the sensory attenuation of self-caused stimuli, where researchers have investigated how these stimuli are perceived as diminished in comparison to externally caused stimuli (Hughes and Waszak, 2011). This is clearly demonstrated in the force-matching paradigm. Here an external force is applied to a subject's finger, after which they must use their other hand to recreate that force as precisely as possible. This takes place under two conditions. In the direct condition, the subject applies force to their finger in a manner as close as possible to pressing on their own finger (given the constraints of the experimental apparatus). In the indirect condition, they apply the force *via* a mechanism located elsewhere, such as a lever to one side. Healthy subjects consistently apply too much force when pressing directly on their finger, indicating that the perceived force is attenuated compared to the other conditions (Pareés et al., 2014). The classical explanation of this effect is that when the brain issues a motor command, it uses a copy of that command to predict the sensory consequences of the resulting motor activity. The predicted sensory input is then subtracted from the actual sensory input, resulting in the attenuation of the stimulus (Klaffehn et al., 2019). This is a representationalist explanation in that it explicitly posits that the brain contains an internal model used to simulate the motor system (Wolpert et al., 1995).

While there is indeed evidence to support the presence of neural correlates of motor activity subsequently influencing sensory perception in different species, specifically *via* corollary discharge circuits (Crapse and Sommer, 2008), the aim of this paper is to interrogate the necessity and utility of internal representations in general and internal predictive models in particular for maintaining adaptive behavior in the presence of self-caused sensory interference. We examine the predict-and-subtract explanation of the sensory attenuation phenomena by using a genetic algorithm (GA) to explore the viable solutions in a dynamical model of a simple embodied system with non-trivial self-caused sensorimotor dynamics, where the task the controller must solve relies on engaging with an environmental stimulus, while its own motor activity also directly stimulates its environmental sensors. Here we focus on the classical, predict-and-subtract approach, which would in theory perfectly solve the interference problem that we have designed, though our GA instead finds alternative, non-predictive solutions which leverage the system's embodiment.

In general, expected stimuli produce a reduced neural response (de Lange et al., 2018). This has been explained in terms of an internal predictive model (e.g., Blakemore et al., 1998, 2000; Wolpert and Flanagan, 2001; Bays et al., 2005; Kilteni and Ehrsson, 2017, 2022; Kilteni et al., 2020; Lalouni et al., 2021). This type of explanation has been described as "cancellation theory," where expected sensations are suppressed (Press et al., 2020). In the interest of completeness, we should mention that there are other predictive accounts of perception, such as Bayesian predictive processing, where attention also plays a major role (Friston, 2009; Clark, 2013; de Lange et al., 2018). The roles of prediction in Bayesian and cancellation theories have been considered contradictory, and "opposing process theory" is one attempt to reconcile them (Press et al., 2020). These alternative approaches are somewhat orthogonal to this project, as they address different potential roles for prediction, whereas we aim to engage with the classical account by investigating the role of embodiment in coping with self-caused sensory interference in a context where prediction and subtraction of that interference is a perfect solution. Likewise, while externally-caused stimuli can also be attenuated, for instance when expected (de Lange et al., 2018), or during movement (Kilteni and Ehrsson, 2022), this paper focuses specifically on coping with self-caused stimuli by modeling a task which requires responsiveness to environmental sensor stimulation despite the presence of self-caused sensory interference.

The problem of ego-noise in robotics hints at why subtracting out self-produced stimuli seems like a natural thing for the brain to do. Ego-noise refers to self caused noise, including that of motors. This noise can interfere with the data collecting sensors of a robot, and the straightforward engineering solution is to cancel out the noise. The explicitly representational and predictive explanation of the sensory attenuation effect meshes well with this engineering perspective, and has informed a predictive approach to dealing with ego-noise (Schillaci et al., 2016). We cite Schillaci et al. here as an illustration that this exact approach has indeed been used in recent work in robotics, and thus our results should have relevance to the field. Of course this is not the only approach to dealing with the general problem of making the self-other distinction in robotics—see for instance Chatila et al. (2018) and Kahl et al. (2022).

In our model, the embodiment is a simple, simulated, two-wheeled system with a pair of light sensors. It is coupled to a controller—a continuous-time, recurrent neural network (CTRNN)—which determines its motor activity. The sensory input to this robot is a linear combination of environmental factors (a function of its position relative to a light) and a self-caused component—a function of the robot's motor activity.

This model is designed to allow both representationalist and non-representationalist solutions to emerge. For the representationalist predict-and-subtract solution to be viable in this model, two criteria need to be met. Firstly, the controller must be able to model the interference. As the controller is a CTRNN, which is a universal approximator of smooth

dynamics (Beer, 2006), it can indeed model the interfering dynamics, which are produced by simple, smooth functions. Secondly, the interference must be able to be removed from the input, given a prediction of the interference. Since the interference is summed with the actual sensor data, it can be removed by subtracting a prediction of the interference from the sensory inputs. This explicitly representational solution would fit with the classical explanation of sensory attenuation. Non-representationalist solutions that take advantage of the system's embodiment are also possible in this model, since the interfering dynamics are a function of the system's motor activity, and are coupled to the controller in a tight sensorimotor loop, embracing the situated, embodied and dynamical (SED) approach. In the classical account, the environmental stimulation of the sensor can be treated as independent of the system's activity, and the self-caused stimulation of the sensor is similarly compartmentalized—the decision to take a particular action is made independently of its incidental sensory consequences, and compensation for these consequences is left to downstream predictive and subtractive processes. In contrast with this approach, modeling how embodied systems are coupled to their environment, in particular how both the system's environmentally and self-caused sensory inputs are influenced by the system's own motor activity, enables additional ways of coping with self-caused stimuli, as will be seen in our results.

Following the evolutionary robotics methodology we explore the space of possible solutions using a genetic algorithm (GA) (Harvey et al., 2005). We then analyze the behavioral strategies of controllers tuned to successfully accomplish a task (phototaxis), in the presence of several different forms of motor-driven sensory interference. This permits us examine a range of ways embodied systems may cope with different self-caused sensory stimuli, and reveals that a number of alternatives to the classical predict-and-subtract approach are viable in our model.

Clearly the simulated robot and neural network controller that we are investigating are very different from humans and their brains. This limits the ability to make direct predictions about humans based on the results found in our model—we don't expect to find people using exactly the same strategies used by the two-wheeled robot. Nevertheless, this type of model can highlight how the solutions found by evolution are not always the same as the solutions that might be identified by a human engineer. As argued by Thompson et al. (1999), humans need to understand what they engineer, to divide and subdivide the problem and solution into smaller units until those units are simple enough to address directly. For example, dividing the problem of coping with self-caused stimuli from the general problems of perception and action, and further dividing it into the prediction and subtraction of self-caused stimuli. Natural or artificial evolution, on the other hand, is under no such constraint. The solutions it finds are the result of iterative improvement with no need for understanding, simplification or compartmentalization. Accordingly, it can find solutions that are "messy" and difficult, perhaps in some cases even impossible, for us to understand. Our evolutionary robotics model, like others before it (Beer, 2003; Phattanasri et al., 2007; Beer and Williams, 2015), allows us to see that there are alternatives to how an engineer might approach solving this particular problem. Furthermore, it allows us to generate concrete examples of alternative strategies for solving the problem at hand, and due to the simplicity of the model these examples are easier to analyze and come to understand than the incredibly complex behavior found in living systems.

In Section 2 we explain the model we developed and the GA we use to optimize its parameters. Then in Section 3 we present the results of our investigation, describing each form of interference used, and explaining the behavior of the most successful system evolved to perform phototaxis in the presence of each form of interference. Finally in Section 4 we summarize the different behaviors evolved to cope with these forms of interference, and discuss how these findings can inform our understanding of the role embodiment plays in coping with self-caused sensory stimuli. We draw attention to how the problem of disentangling self-caused and environmental stimulation of the sensors is made easier for embodied systems by the influence embodied systems have over both self-caused and environmental stimulation of their sensors, and we argue that, for embodied systems, this problem need not require the use of an internal model.

## 2. Model and methods

In this section we first describe our model of an embodied system with self-caused, motor-driven sensory interference, which must perform a task where clear perception of the environment is beneficial. We then describe the genetic algorithm (GA) that we use to investigate how embodied systems can cope with self-caused sensory input.

### 2.1. Model

We model a simple light-sensing robot, controlled by a neural network, where the robot's light sensors can also be directly stimulated by the robot's own motor activity. The two-wheeled robot moves about an infinite, flat plane. It has a pair of directional light sensors, and the environment contains a single light source. Over the course of a single simulation, this light source's position remains fixed. The robot is controlled by a continuous-time, recurrent neural network (CTRNN). Motor-driven interference is ipsilateral and non-saturating, and is determined by one of three different functions, which are detailed in the Experiments section. Figure 1 provides a visual overview of the model architecture. As the model is fully deterministic, the course of each simulation is fully determined

by the robot's initial distance from and orientation toward the light. In each simulation, the robot begins at the origin $(0, 0)$, facing toward positive $y$, and initial conditions are varied by positioning the light at a different $(x, y)$ coordinate.

### 2.1.1. Embodiment

The robot is circular, with two idealized wheels situated on its perimeter $\pi$ radians apart, at $-\pi/2$ and $\pi/2$ relative to its facing. The wheels can be independently driven forwards or backwards. Its two light sensors are located on its perimeter at $-\pi/3$ and $\pi/3$ relative to its facing. The environment it inhabits is defined entirely by the spatial coordinates of the single light source. The robot's movement in its environment is described by the following set of equations:

$$\dot{x} = (m_L + m_R)\cos(\alpha) \tag{1}$$

$$\dot{y} = (m_L + m_R)\sin(\alpha) \tag{2}$$

$$\dot{\alpha} = (m_R - m_L)r \tag{3}$$

Where $x$ and $y$ are the robot's spatial coordinates, and $\alpha$ is the robot's facing in radians. $m_L$ and $m_R$ are the robot's left and right motor activation, respectively, and are always in the range $[-1, 1]$. The values of $m_L$ and $m_R$ are specified by the controller, which is described later. $r = 0.25$ is the robot's radius. We simulate this system using Euler integration with $\Delta_t = 0.01$.

Physically this describes positive motor activation turning its respective wheel forwards, and conversely for negative motor activation. If the sum of the two motors' activation is positive, the robot as a whole moves forwards with respect to its facing, while if it is negative, the robot moves backwards. The amount that the robot turns is also determined by the relationship between the two wheels.

The robot's two light sensors are located at the coordinates $(x + cos(\alpha + \theta)r, y + sin(\alpha + \theta)r)$, where $\theta$ is the sensor's angular offset. For the left sensor, $\theta = \pi/3$ and for the right sensor, $\theta = -\pi/3$. The environmental stimulation of the sensors is given by:

$$s = \frac{(\boldsymbol{b} \cdot \hat{\boldsymbol{c}})^+}{1 + D^2}\boldsymbol{\epsilon} \tag{4}$$

Where $\boldsymbol{b} = [cos(\alpha + \theta), sin(\alpha + \theta)]$ is the unit vector pointing in the direction the sensor is facing, and $\boldsymbol{c}$ is the vector from the sensor to the light, with $\hat{\boldsymbol{c}}$ denoting that the vector is normalized to have a unit length. That is $\hat{\boldsymbol{c}} = \boldsymbol{c}/|\boldsymbol{c}|$, where $|\boldsymbol{c}|$ is the magnitude of $\boldsymbol{c}$. The symbol $\cdot$ denotes the dot product of the two vectors, and the superscript $^+$ indicates that any negative values are replaced with 0. $D$ is the Euclidean distance from the sensor to the light, and $\boldsymbol{\epsilon} = 5$ is a fixed environmental intensity factor. $s_L$ denotes the activation of the left sensor, with $\theta = \pi/3$, while $s_R$ denotes the activation of the right sensor, $\theta = -\pi/3$.

The numerator is maximized at 1 when the sensor is directly facing the light, and minimized at 0 when the sensor is facing $\pi/2$ radians (90°) or more away from the light. The denominator is minimized at 1 when the distance from the sensor to the light is 0. This means that the activation of a sensor grows both as the sensor faces more toward the light, and as the sensor approaches the light (so long as it is facing less than $\pi/2$ radians away from the light).

### 2.1.2. Controller

The controller is a continuous-time recurrent neural network (CTRNN) defined by the state equation below, following Beer (1996):

$$\tau_i \dot{y}_i = -y_i + \sum_{j=1}^{N} \omega_{ji}\sigma(y_j + \beta_j) + I_i \tag{5}$$

Here $N = 10$ denotes the number of neurons in the network. $y_i$ indicates the activation of the $i$th neuron. The parameter $\tau_i$ is the time constant of that neuron, where $0 < \tau_i < 3$, while the parameter $\beta_i$ is its bias, where $-5 < \beta_i < 5$. $I_i$ is any external input to the neuron. $\sigma(x) = 1/(1 + e^{-x})$ is the standard logistic activation function for neural networks, and is a sigmoid function in the range $[0, 1]$. $\omega_{ji}$ is a weight determining the influence of the $j$th neuron on the $i$th neuron, where $-5 < \omega_{ji} < 5$.

Two neurons are designated as input neurons, and all their incoming interneuron weights $\omega_{ji}$ are set to 0, including the recurrent weight $\omega_{ii}$. With the robot described above, neurons 1 and 2 are designated as input neurons, and $I_1 = w_I s_L$, while $I_2 = w_I s_R$, where $w_I = 5$ is a fixed input scaling weight. These are the only neurons which receive an external input, so $I_{3..N} = 0$ always.

Two neurons are designated as output neurons (neurons 9 and 10), and their activation values $y$ are treated as the output of the network. In our case, $y_{N-1}$ and $y_N$ provide the values $m_L$ and $m_R$, respectively. Output is scaled to be in the range $[-1, 1]$ by the function:

$$o(y) = \frac{2}{1 + \exp(\frac{-y}{\sqrt{\omega_{\max}}})} - 1 \tag{6}$$

Where $\omega_{\max} = 5$ denotes the maximum weight value $\omega$ permitted for a node in this CTRNN. The two output neurons do not receive stimulus from the input neurons. That is if $j \in \{1, 2\}$ and $i \in \{9, 10\}$ then $\omega_{ji} = 0$. The remaining six neurons are interneurons, each of which receives inputs from all other neurons in the network. This neural network architecture is illustrated in Figure 1.

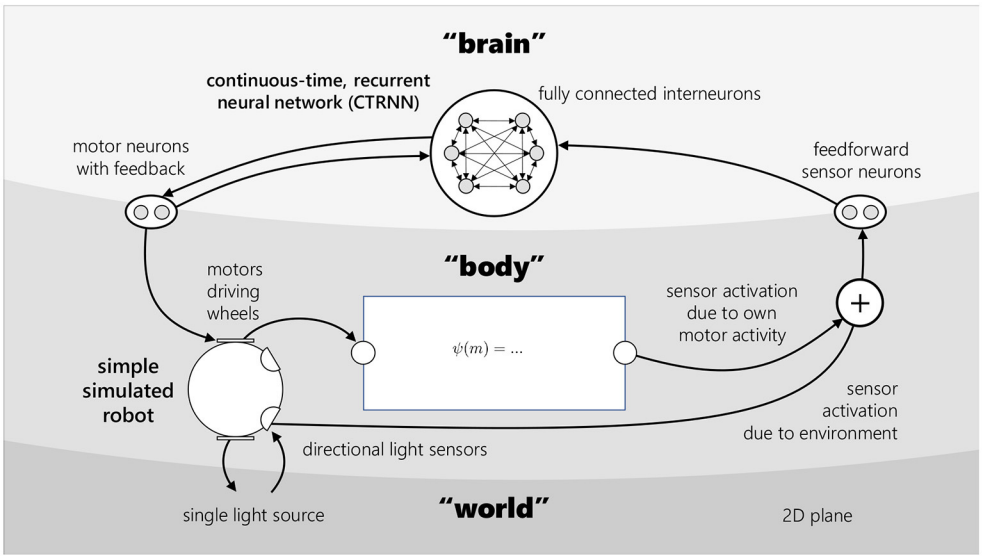



FIGURE 1
An embodied model with motor-driven sensory interference. This model is used throughout the paper. It consists of three parts—the "brain," the "body" and the "world." The brain is a continuous-time, recurrent neural network (CTRNN), with 6 fully connected interneurons, 2 sensor neurons which project to all interneurons, and 2 motor neurons which project to and receive projections from all interneurons. The motor neurons determine the activation of the body's 2 motors. The body's position and orientation relative to the single light source in the environment determine the activation of its 2 light sensors. The value received at a given point in time by the right sensor neuron is a linear combination of the right light sensor activation, and a function $\psi$ of the right motor's activation, representing self-caused sensory stimulation—and likewise for the left sensor, sensor neuron, and motor.

### 2.1.3. Motor-driven interference

Perception necessarily involves both the system and its environment. Nevertheless, we can consider the degree to which the activity of the system or environment contributes to a given stimulus. Let us take three very different points in this space. (1) If our robot passively sat still, while a light in the environment turned on and off, the change in the light sensors' activations would primarily be due to external causes—the robot's own activity would not play a role. (2) On the other hand, in the model described above, all changes in the light sensors' activations are the result of a change in the *relationship* between the light's position and the robot's position and facing. Because the light is static, the change is induced by the robot's activity, but determined by the robot's spatial relationship with its environment. (3) At the other end of the scale from (1), consider the case where the robot inhabits a lightless environment in which its sensors are directly and exclusively stimulated by its own motor activity. In this case, neither external causes, nor the relationship between the system and the environment play a role—the change in the sensors' activation is due solely to the robot's own activity.

For living systems in the real world, none of these three points are typically possible—for (1) perception is rarely (if ever) purely passive, for (2) movement will likely involve self-produced sensations even if the environment is passive, and for (3) self-produced sensations will depend on environmental conditions. Nevertheless, our own experiences may lie closer to one of these points than to another. Consider the visual experience of (1) sitting watching a movie (a passive experience, yet one whose visual sensations will still depend on activities like movement or blinking), (2) turning to look around the otherwise still room briefly (where the visual stimulation is largely determined by the spatial relationship between the eyes and the room, but still influenced by changes in the environment like the ongoing movie, and self-produced sensations like the peripheral vision of bodily movement), then (3) scratching your nose (where a change in visual stimulation is caused by your own hand entering the visual field, but depends also on static and dynamic environmental factors like the general lighting of the room and the flickering light of the movie screen).

In the model described so far, there is no possibility for directly self-caused stimuli like (3). This is precisely the kind of self-caused sensory input we are concerned with here, so we extend the model with an interference function $\psi(m)$. The various interference functions we study are described in Section (3). The interference function is used in a new sensory input equation:

$$s' = \lambda\psi(m) + (1 - \lambda)s \tag{7}$$

Where $s$ is the original light sensor activation, $m$ is the ipsilateral motor's output, and $\lambda$ is a scaling term controlling how much of the sensory input is due to the environment, and how

much is due to the system's motor activity. Substituting for the original input neuron equations, this gives:

$$I_1 = \omega_I s'_L = \omega_I(\lambda\psi(m_L) + (1-\lambda)s_L) \quad (8)$$

$$I_2 = \omega_I s'_R = \omega_I(\lambda\psi(m_R) + (1-\lambda)s_R) \quad (9)$$

This combination of motor-driven interference with sensor activity is additive and non-saturating. That is, the interference $\psi(m)$ can never be so high that change in the environmental stimulation $s$ of the sensor does not result in a change in $s'$. This means that if $\psi(m)$ can be predicted by the network, then this value can simply be subtracted from the input neuron's output to other nodes. This mapping also uses the ipsilateral motor to generate interference for each sensor. This was chosen for two reasons. Firstly, it is physically intuitive. Secondly, because the motor neurons have recurrent connections to the interneurons, this means that the neural activity determining $m_L$ and $m_R$ [and thus $\psi(m_L)$ and $\psi(m_R)$] contributes to the interneurons' synaptic inputs, making prediction easier.

To summarize, we start with a model of a two-wheeled robot with two light sensors, controlled by a CTRNN. In this model, changes in a light sensor's activation are purely the result of the robot's position and orientation changing relative to the light. We extend this model by adding a function which, given a motor activation value, produces an interfering output. Instead of the input neurons of the controller directly receiving the current activation of the light sensor, the light sensor's activation is first combined with this interference. The parameter $\lambda$ controls the weighting given to the sensor activation vs. the interference in this combined term. For example, with $\lambda = 0.05$, instead of the light sensor's true reading $s$, the controller receives $0.95s + 0.05\psi(m)$. The interference functions $\psi(m)$ are described in the Section 3.

## 2.2. Methods

Parameters for the CTRNN controller were evolved using a tournament based genetic algorithm (GA) based on the microbial GA (Harvey, 2011). The GA operates on a population, which consists of a number of solutions specifying the parameters for the CTRNN. In a tournament, two randomly chosen solutions from the population are evaluated independently. Their fitness is compared, and then in the reproduction step the lower scoring solution is removed from the population and replaced by a mutated copy of the higher scoring solution. Our microbial GA differs from the classic presentation in that it ensures that each member of the population participates in exactly one tournament before the reproduction step is performed for the entire population. This allows generations of the population to easily be counted.

The following parameters were evolved for each node $i$ in the CTRNN: the time factor $\tau_i$, the bias $\beta_i$, and a weight vector specifying the incoming interneural weights for node $i$, where $\omega_{ji}$ refers to the weight applied to the connection from $j$ to $i$.

Each evolvable parameter of the network is encoded in the genome as a single 32 bit floating point number in the range [0, 1]. The weights and biases are translated from gene $g$ to phenotype $\omega$ or $\beta$ *via* the linear scaling function $(\omega_{min} + \omega_{max})g + \omega_{min}$, where $\omega_{min}$ and $\omega_{max}$ are the minimum and maximum neural weights, respectively –5 and 5, while for $\tau$ we use the exponential mapping $e^{3g}/10$.

The reproduction procedure used, based on the result of a tournament, is to remove the loser from the population, and add in its place a copy of the winning genome. Each gene in this copy is then mutated by the function

$$m(g) = ((g + X\mu) + 1) \bmod 1 \quad (10)$$

Where $X \sim \mathcal{N}(0,1)$ is a random variable drawn from a normal distribution with a mean of 0 and a standard deviation of 1, $\mu = 0.2$ is the mutation factor, and the result is scaled by adding 1 and taking the modulo with 1 to ensure the result is in the range [0, 1].

In all cases the system was evolved to perform phototaxis using the following fitness function:

$$\frac{\sum_{t=0}^{T} d(x_t, y_t)^2 t}{\sum_{t=0}^{T} t} \quad (11)$$

Where $t$ is the time at the current integration step, $T$ is the trial duration, and $d(x, y)$ is the euclidean distance from the point $(x, y)$ to the light. The squared distance is used rather then the actual distance here solely for computational efficiency. Multiplying the distance by the current time means that minimizing distance later in the trial is more important to the fitness score than doing so earlier is. The final distance is the most important, while the original distance from the light at $t = 0$ is completely disregarded. However, improvement at any time is always relevant: $t = 99$ is almost as important as $t = 100$.

In each trial, the robot begins at the origin. Each generation, four light coordinates are stochastically generated. The first coordinate is chosen uniformly at random to lie on a circle of radius 3 centered on the origin. The other three coordinates lie on the same circle and form a square with the first. Each solution in the population has its fitness score calculated for each of the four light coordinates. These scores are combined before comparison in the tournament. This means that a given solution's score may go up or down from generation to generation, as it may perform better or worse on that generation's set of light coordinates. This helps prevent the GA becoming stuck in a local optima.

A population of 50 individuals was used. The trial duration was chosen to allow enough time for robust phototaxis to be selected for, either 10 or 20 time units depending on the interference function. The GA was allowed to run for a sufficient

number of generations for fitness gains to plateau and for the population of solutions to converge.

## 3. Experiments

To investigate how embodied systems cope with motor-driven interference, we began by using the GA to find parameters that would allow a CTRNN controller to perform phototaxis in the basic model with $\lambda = 0$ (i.e., with no motor-driven sensor interference). The population of controllers that were the product of this GA run are taken as the *ancestral population* for the subsequently evolved populations in Experiments 2–4. That is, parameters for these populations were evolved starting from this ancestral population, rather than starting from a new, random population. We chose to use an ancestral population, rather than evolving subsequent populations from scratch, in order to allow for direct comparison between the behavior of the systems optimized with and without the presence of motor-driven interference. The results of Experiment 1 are presented in Section 3.1.

In addition to Experiment 1 with the basic version of the model where $\lambda = 0$ (and therefore $s' = s$), we consider three further versions of the model in Experiments 2–4, each corresponding to a different interference function $\psi(m)$. We use $\lambda = 0.5$ with each of these three functions. In turn we consider: (i) a threshold-like sigmoidal function, whose interference can be completely avoided by appropriately modified behavior; (ii) a form of unavoidable interference, taking the square of the motor activity; and (iii) a time-dependent interference function, a sine wave whose frequency depends on the motor activity, which eliminates a degree of control that was present with the squared interference. The three interference functions used for these experiments can be seen in Figure 2, and are introduced and explained in more depth in Sections 3.2–3.4, where the corresponding results are also presented.

### 3.1. Experiment 1: Phototaxis without interference

A highly fit population of controllers was evolved to perform phototaxis in the basic model, with no motor-driven interference. Evolution of this population began from a population of solutions with uniformly random interneuron weight and time constant values, and with center-crossing biases (Mathayomchan and Beer, 2002). A trial duration of 10 time units was used. After evolution, genomes for this population are highly convergent, indicating that the population has become dominated by a single solution. Examining the fittest member of this population, we found that the controller reliably brought the robot close to the light across a collection of light coordinates representative of those used during evolution (Figure 3). The robot's behavior results in it remaining close to the light even over time periods orders of magnitude longer than the trial duration used during evolution. This indicates that the solution produces a long term, stable relationship with the environmental stimulus.

The ancestral solution's behavior is well preserved in the descendent populations evolved to handle the various interference functions studied. Understanding how this solution works is helpful for understanding how the descendent solutions handle motor-driven sensory interference.

The ancestral solution's behavior can be divided into 2 phases:

A) The *approach* phase, where the robot makes its way close to the light. This phase has to account for the light starting at an unknown point relative to the robot.
B) The *orbit* phase, where the robot's long-term periodic activity maintains a close position to the light.

Note that this two phase description does not imply switching between two different sets of internal rules. These phases are driven by the ongoing relationship between the robot and its environment, and are better thought of in dynamical systems terms as a transient and a periodic attractor.

The orbit phase (Phase B) is simpler to explain, so we will begin with it. Here we can approximate the robot's behavior with a simple program:

1. Approach the light while driving backwards, such that you will pass the light with the light on your right hand side.
2. When the light abruptly enters your field of vision, it causes a spike in your right sensor: quickly respond by switching to driving forwards instead, turning gently to the left.
3. After driving forward has brought the light behind you and out of the sensor's field, go to 1.

We observed this behavior across all the light coordinates we examined. Figure 4 and the corresponding caption explains how this behavior applies to the trajectory for a specific light coordinate, showing how the simple program described above matches its behavior. The left sensor is completely uninvolved in this process. In fact for some initial light positions, namely when the robot begins with the light on its right, the left sensor is also completely uninvolved in the approach phase. That is, if the left sensor is completely deactivated throughout certain trials, the trajectory is completely identical to if it were active.

The approach phase (Phase A) often consists of simply driving forwards, and then continuing to drive forwards until the right sensor is not stimulated. Thereafter the procedure for Phase B is followed, with the approach differing from the orbit primarily in that the amount of time spent on each step of the 'program' while approaching the light varies more than it does when the robot is stably orbiting the light. This is what we see in trajectory shown in Figure 4, and in all conditions when the left sensor is not stimulated. However in conditions when the left

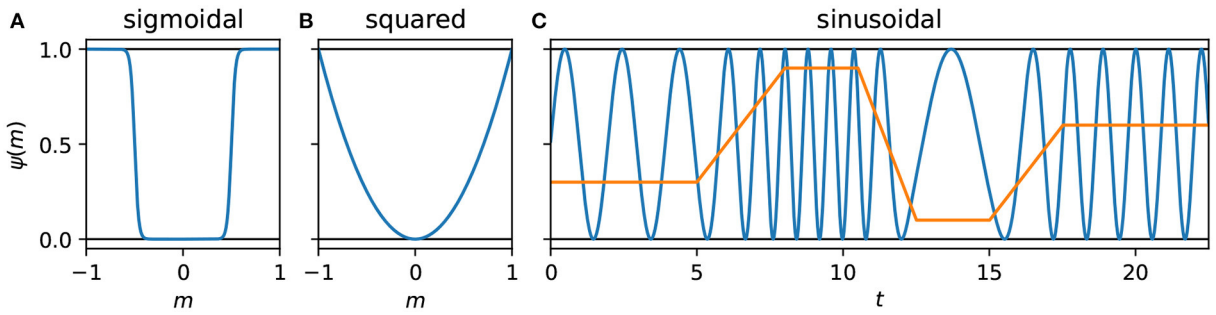



FIGURE 2
Plots of the interference functions used in Experiments 2–4. **(A,B)** Plot pure functions of $m$ corresponding to Equations (12) and (13) (Experiments 2 and 3, respectively). **(C)** Plots a function of time that depends on the cumulative history of $m$, Equation (14) (Experiment 4). The blue line is the interference, while the orange line is the motor activity.

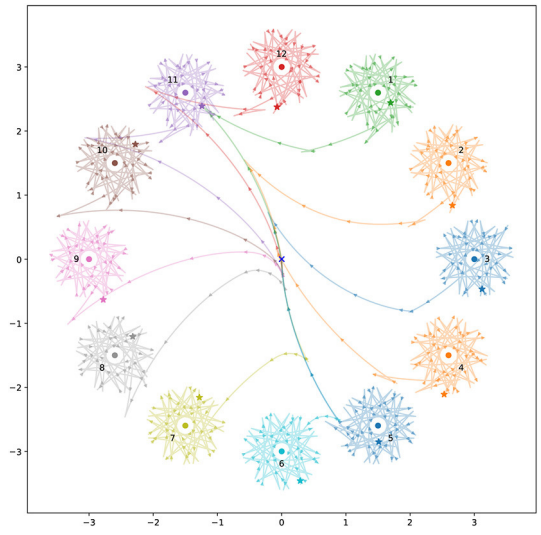


FIGURE 3
Spatial trajectories for the best individual from the ancestral population for 12 different light coordinates. The robot always begins at the origin, facing toward positive $y$ (upwards). Stars mark the final position reached during the trial duration used during evolution. The colored circles show the light position for the correspondingly colored trajectory. The triangles along the trajectories point in the direction the robot is facing. They are plotted at uniform time intervals, so more spaced out triangles indicate faster movement.

sensor is stimulated during the approach phase, the left sensor is involved in guiding the robot into a state where Phase B takes over. This can be seen in Figure 5.

This solution is an instance of a more general robust strategy for performing phototaxis in this model, which can be summarized even more simply as:

- If you don't see the light, drive backwards (it must be behind you).
- If you do see the light, drive forwards until you can't see it any longer.

The reason this does not result in just driving backwards and forwards along the same arc is that the robot turns a different amount when driving forwards vs. when driving backwards. The turn amount is determined by $m_R - m_L$, while the direction of travel is determined by whether $m_R + m_L$ is negative or positive. When adjusting motor activity to change directions, it's trivial to also change the amount of turn. Of course this general strategy is not a complete description of the robot's behavior, the effect of sensor stimulation can be time dependent and differ for the left and right sensors. Particularly during Phase A, the approach to the light, the exact trajectories taken by the robot depend on continually regulating the 2 independent motors' speed and direction of activity to perform both gradual turns and sharp changes in direction *via* 3 point turns with sufficient precision to reliably enter Phase B and maintain it. However, we see this general strategy well preserved in populations descendent from this ancestral population as well as evolved independently in non-descendent populations.

To summarize, the ancestral solution takes advantage of the particular nature of its sensors, driving backwards so that the sensors are stimulated sharply. It adjusts its motor activity in response to this sharp stimulation in such a way that the stimulation is extinguished. This environmentally mediated negative feedback loop plays a critical role in enabling the system to remain stably in close proximity to the light source. Capturing this type of natural feedback loop is a strength of modeling work following the SED approach. In the subsequent sections, we will see the role this pattern of behavior plays in coping with additional self-caused interference, and how this behavior is modified when this population of solutions is taken as the ancestral population for subsequent optimisation *via* the GA with the addition of motor-driven interference.

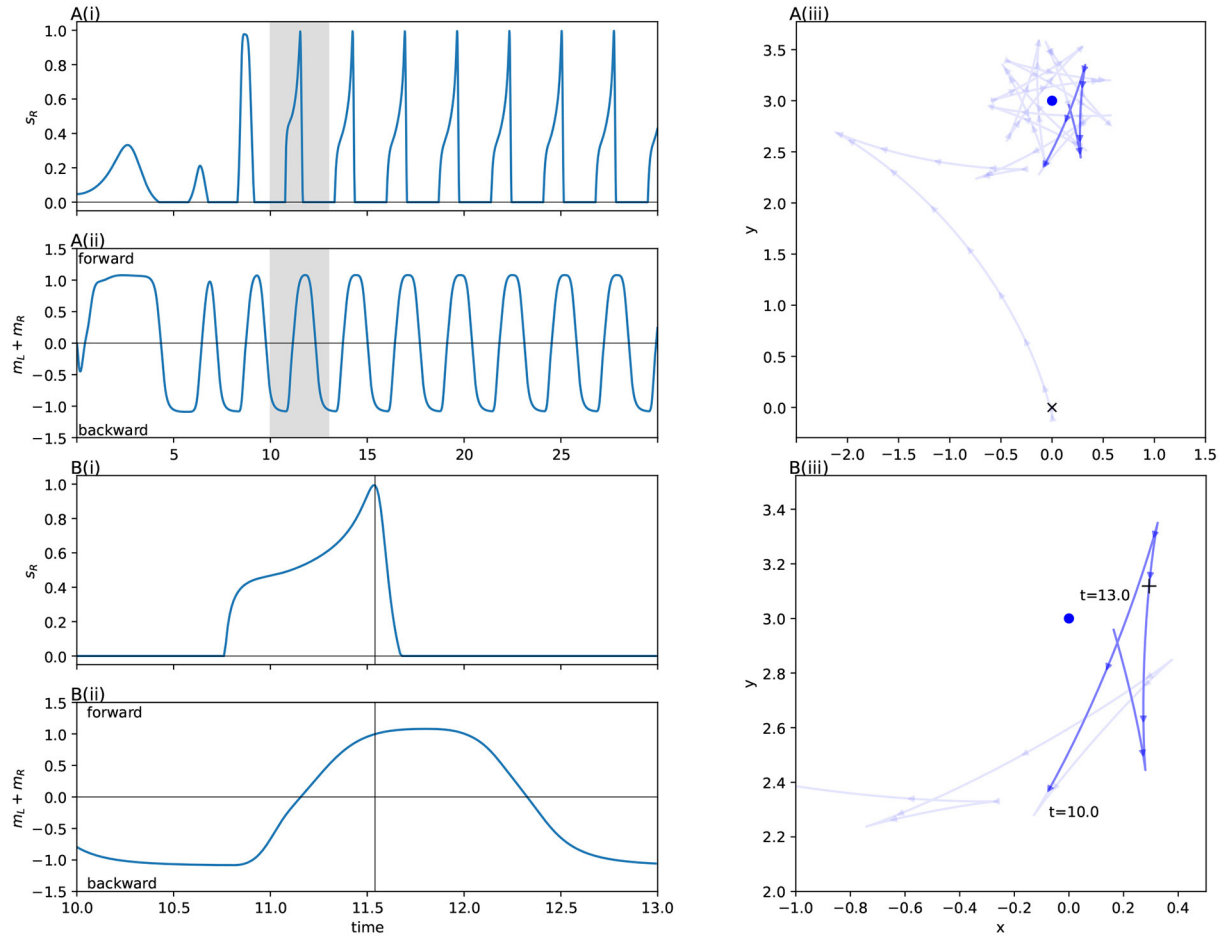



FIGURE 4
Detail of the orbit phase (Phase B) for the ancestral solution. The plots marked **(A)** Show the ancestral solution when the light is at coordinates (0, 3)—position 12 in Figure 3. The highlighted sections of these figures mark the time period 10–13, which is shown in more detail in **(B)**. The vertical line in **B(i,ii)** marks the peak of right sensor activation, which corresponds to the + in **B(iii)**. The activity shown in **(B)** corresponds to the Phase B program (see main text). Before $t = 11$, the robot drives backwards, passing the light on its right side. As the right sensor is stimulated, the robot changes direction, driving forwards. After the right sensor stimulation peaks and dies down, the robot changes direction again, reversing toward the light. **(A)** Show how the process repeats.

## 3.2. Experiment 2: Avoidable interference

Having evolved a system to perform phototaxis in the absence of directly self-caused sensory stimuli, we take this population of solutions as the ancestral population for subsequent evolution in the presence of motor-driven interference functions to begin investigating how embodied systems can cope with this type of interference. In this section we describe the first form of self-caused sensory interference modeled, and how the ancestral solution is modified to accommodate it.

The simplest possible interference would be adding a constant value to all the sensor inputs. However this would not depend on the system's motor activity. Therefore the first $\psi(m)$ that we model is a threshold-like interference function, where interference is maximized when motor activation is above a threshold value, and $\approx 0$ elsewhere. To achieve this effect with a smooth function, we use a relatively steep sigmoidal function, with the equation:

$$\psi(m) = \frac{1}{1 + \exp(-k(|m| - p))} \tag{12}$$

Where $\exp(x) = e^x$ and $|m|$ is the absolute value of $m$, and where $k = 50$ is the term controlling the steepness of the sigmoid's transition from 0 to 1, while $p = 0.5$ determines the midpoint of the transition. So when $m < -0.5$ or $m > 0.5$: $\psi(m) \approx 1$ and when $-0.5 < m < 0.5$: $\psi(m) \approx 0$. This function is unique among the three in that were the system to constrain

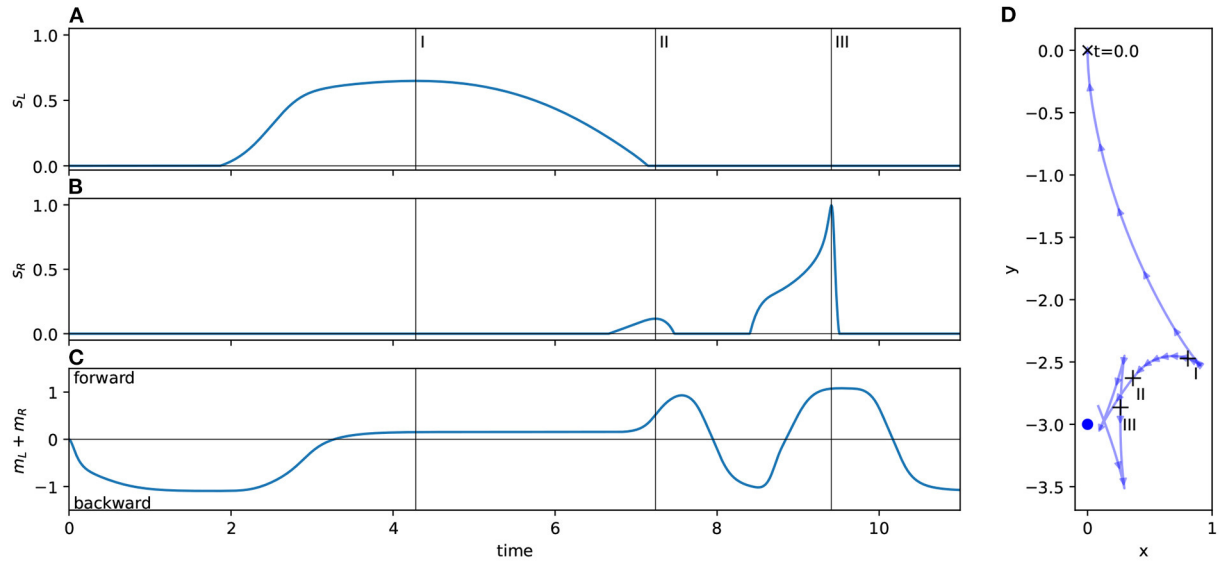



FIGURE 5
An example approach phase (Phase A) for the ancestral solution which is guided by the left sensor. **(A–C)** Show the sensorimotor activity of the ancestral solution when the light is at coordinates (0, −3), position 6 in Figure 3. **(D)** Plots the spatial trajectory of the robot. The vertical lines in plots **(A–C)** show the peaks in sensor activity. These correspond to the + markers in **(D)**. Initially the robot drives backwards. The left sensor stimulation between $t = 2$ and $t = 8$ is associated with the robot to driving forwards while turning strongly to the left. Once this turn has oriented the robot such that the right sensor is being stimulated and the left sensor is no longer being stimulated, the robot drives forward until the right sensor is no longer stimulated. From here, this is just the same Phase B behavior presented in Figure 4.

its motor activity to the appropriate range, it would avoid the interference altogether. We will refer to the interference generated by this function as *avoidable* or *sigmoidal* interference.

With motor activity capped at 50%, motor-driven interference can be avoided, and phototaxis can still be performed, just more slowly. Moving more slowly comes at a cost to fitness though, since the fitness function (Equation 11) rewards reaching the light quickly. Therefore, a predict-and-subtract solution to the interference which preserves the speed of the high-performance ancestral solution should outperform a solution which simply avoids the interference. However, we instead found that the fittest solution from the 5 populations evolved to perform phototaxis with the sigmoidal interference function modifies the motor activity of the ancestral solution significantly.

Figure 6 illustrates how the characteristic motor activity of the solution evolved with sigmoidal interference differs from that of the ancestral solution. Keeping in mind that the ancestral solution often involved minimal environmental stimulation of the left sensor, we observe that the left motor in this evolved solution never produces interference. This comes at the cost of greatly decreased absolute motor activity relative to the ancestral solution. The ancestral solution's left motor activity ranges widely, from −0.96 to 0.10 with a median of −0.82, close to the maximum possible absolute value of 1. See Figure 4A(ii) for ancestral motor activity as a time series. In contrast, the left motor activity of this solution ranges only between −0.42 and −0.32 with a median value of −0.38. Time series of this motor activity can be seen in Figures 8A(iv),B(iv). This drastic decrease in motor activity lowers the speeds attainable by the robot, but prevents motor-driven interference with the left sensor. While the activity of the left motor is kept below the threshold for producing interference at all times, keeping the left sensor free of interference, the right motor does produce interference. The distribution of right motor activity is bimodal, with peaks just below the interference threshold of 0.5, and close to its maximum value of 0.84. This bimodal distribution is the result of this solution producing two distinctly different orbit types.

The orbiting behaviors of this system are of interest because they demonstrate ways in which a long term, stable relationship with an environmental source of sensor stimulation can be maintained in a model with motor-driven sensor interference. As with the ancestral population, a trial duration of 10 time units was used for this population. Due to the decreased overall motor activation relative to the ancestor, and the consequently decreased speed, the robot does not get as close to the light in that time as the ancestor did. This means that what has been selected for by the genetic algorithm here is modification of the approach phase to maintain accuracy in the presence of this novel interference. However, due to a sufficiently accurate approach and the evolved regulation of the motor-driven interference, stable orbits are still achieved across all light

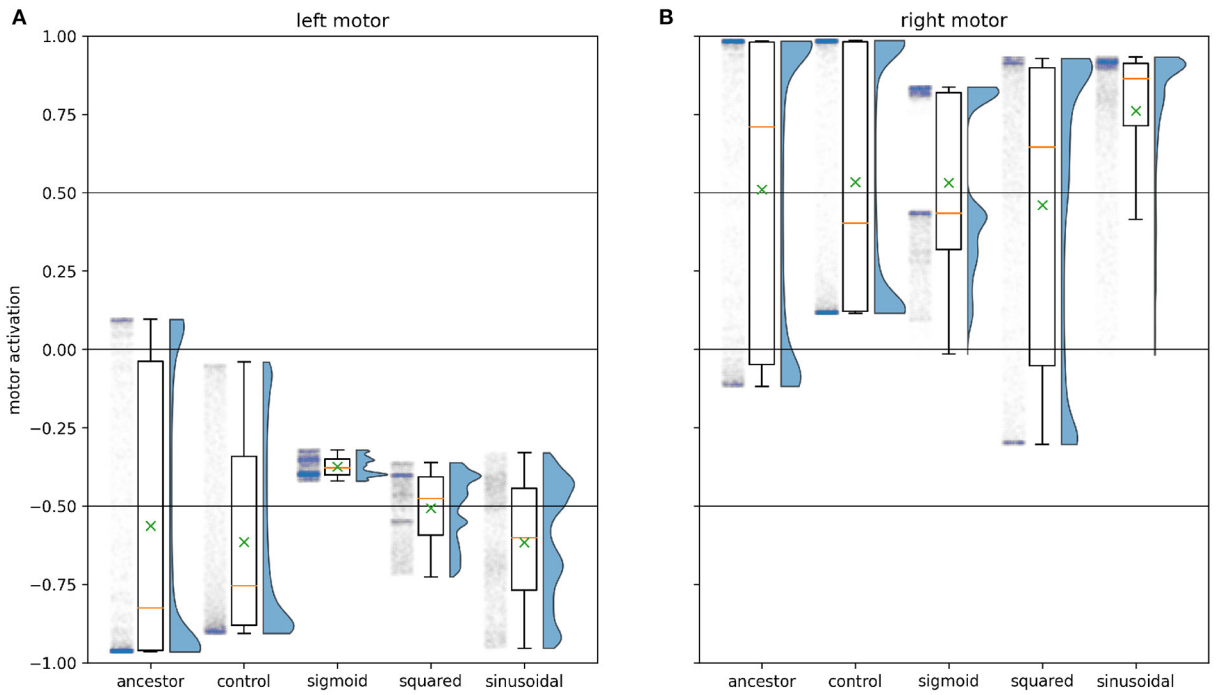



FIGURE 6
Motor activity for the 12 light positions shown in Figure 3 for time 20 to time 50 (integration steps 2,000–5,000), for the evolved solution to each experiment. This is one way of visualizing aspects of the ancestral behavior that have (and have not) been modified by further evolution in the presence of an interference function. The boxes extend from the first to the third quartile of the motor activity, and contain a yellow line showing the median, and a green × showing the mean. The whiskers extend to 1.5 times the inter-quartile range. The half-violin plot to the right of each box plot estimates the distribution of the motor activity, while to the left is a scatter plot of each simulated moment of motor activity with randomized horizontal placement. The column labeled control plots exactly the same information for the fittest solution evolved with $\lambda = 0.5$ and the null interference function $\psi(m) = 0$, showing the scope of change seen simply due to the presence of $\lambda$ and to genetic drift. **(A)** plots the motor activity for the left motor of each system, while **(B)** plots this information for the right motor. Of particular relevance to the solutions cataloged in this paper are the depressed (absolute) left motor activation with sigmoidal interference and the corresponding bimodal distribution of right motor activity; the reduced range of left motor activity with squared interference, and the fact that the right motor activation with squared interference continues to cover a wide range; and the reduction in low (absolute) values of motor activity with the sinusoidal interference function.

positions in the very long term. Unlike the ancestor, we see two distinctly different orbit behaviors. Across all interference functions we refer to those orbits reminiscent of the ancestral solution, involving forward and backward motion around the light, as Type 1 orbits, and to orbits which loosely circle the light while driving forwards as Type 2 orbits. These are easily distinguished visually (see Figure 7). As with the ancestor, approaches can broadly be divided into those guided by the left sensor, and those that are not. In the majority of cases for this solution, the approach phase preceding Type 1 orbits is guided exclusively by the right sensor, while Type 2 orbits tend to follow a left sensor guided approach phase.

Type 1 orbits come much closer to the light. They display similar sensorimotor behavior to the ancestor's orbit behavior (Phase B), maintaining a stable relationship to the light by repeatedly driving backwards and forwards, albeit with greatly reduced motor activity compared to the ancestor. Figure 8A shows a typical example of sensorimotor activity for Type 1 orbits. Right motor-sensor interference is almost entirely avoided. A very low amount (not visible in the figure) coincides with the robot driving forwards slowly. This interference is necessary because the left motor's activity is negative, and is maintained very closely to the threshold for interference, so the right motor's positive activity cannot be raised sufficiently highly to drive forwards without producing at least a small amount of interference. We summarize this orbit strategy as performing the known good ancestral strategy while constraining motor activity to avoid sensor interference.

Type 2 orbits loosely circle the light, and are very different from the ancestral orbit behavior. Figure 8B shows an example of typical sensorimotor activity for this type of orbit. These orbits do not involve environmental stimulation of the right sensor, instead the left sensor is stimulated throughout the orbit phase. Unlike Type 1 orbits, where the relationship to the light is maintained by repeatedly driving forwards and backwards, the robot exclusively drives forwards. It does so very quickly, producing high right motor-sensor interference. We characterize this orbit strategy as keeping "one eye on

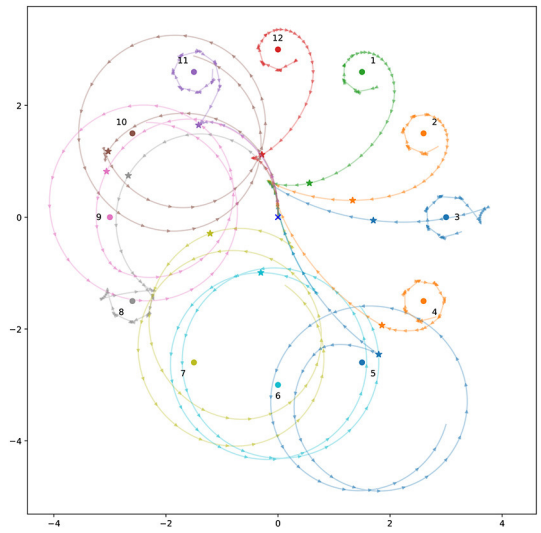



FIGURE 7
Two distinct types of orbits are visible in the spatial trajectories for the best individual from populations evolved with sigmoidal interference (Equation 12). Type 1 orbits, reminiscent of the ancestral solution, are seen for Lights 11, 12, 1, 2, 3, and 4. Type 2 orbits, which feature a forward moving, counter-clockwise orbit of the light are seen for Lights 5, 6, 7, 9, and 10. For Light 8, an approach typical of a Type 2 orbit instead puts the robot in position for a Type 1 orbit.

the prize," where the left sensor, facing the light, is kept free of interference. Meanwhile the right sensor, facing away, is continually stimulated by the right motor's activity. This orbit strategy is uniquely enabled by the ipsilateral nature of the motor-driven sensory interference.

In the presence of this threshold based interference, the best solution found by our GA when modifying the ancestral population to accommodate this interference constrains the ancestral solution's motor activity to avoid interference while performing the same function of phototaxis, using (in some situations) the same basic strategy. This approach contrasts with the predict-and-subtract approach of modifying the controller to subtract the anticipated interference from the sensor neurons' outputs, allowing the behavior of the ancestral solution to be performed without modification. This suggests that in our model such solutions are far closer in evolutionary space to the ancestral solution than a predict-and-subtract solution would be. The relevance of this to the evolutionary history of biological control systems is unclear, however it may suggest that adjusting neural activity to accommodate a novel form of motor-driven sensory interference would involve regulation of the behavior producing that interference in addition to or instead of the neural subtraction of internally predicted interference. This demonstrates that behavior modification does indeed work as a solution to motor-driven sensory interference, and that the precise way in which behavior is modified can depend heavily on the particularities of the sensorimotor contingency in question. Specifically we have seen how two ways of compensating for motor-driven sensory interference emerged in our model. Firstly, motor activity may be constrained to ranges that minimize or avoid interference with the sensors. Secondly, interference can be avoided for only one sensor, which is kept trained on relevant environmental stimuli. This permits unconstrained use of motor activity which interferes with the other sensor. While this robot is clearly much simpler than a human, this demonstration of how pre-existing behavior can be modified to avoid the effects of novel, self-produced sensory interference may suggest a role for such solutions in other contexts, such as less complex organisms (including perhaps our deep evolutionary past) and simple robots.

## 3.3. Experiment 3: Unavoidable interference

Sigmoidal interference certainly does not exhaust the possibilities for modeling interference, nor does it capture the fact that many self-caused stimuli cannot be avoided when taking action. Therefore, we also model non-avoidable interference, where the interference increases with the absolute magnitude of the motor activation. To minimize discontinuities in the system, and to ensure the interference can be approximated by the CTRNN controller, we use a smooth function—the square of the motor activity:

$$\psi(m) = m^2 \tag{13}$$

We will refer to the interference generated by Equation (13) as *unavoidable* or *squared* interference. Like the avoidable, sigmoidal interference function modeled previously, the magnitude of the interference correlates with the magnitude of the motor activity. However, unlike with the avoidable interference function, now all changes in motor activity produce a corresponding change in the sensory interference.

Examining the fittest solution produced by the GA's modification of the ancestral solution, we again find the ancestral solution well preserved. A trial duration of 20 time units was used during evolution to compensate for any decreased speed compared to the ancestor. The general strategy of approaching the light while driving backwards is maintained, however motor activity has changed to accommodate the addition of the squared interference function. The left motor's activity is now constrained to a much smaller range (see Figure 6A), which lowers interference dramatically compared to the interference that would be produced by the ancestral solution's motor activity (see Figure 9A). The right motor generates significant interference, but we find that rather than destructively interfering with the sensor in such a way that the environmental stimulus is masked, this motor-driven sensor

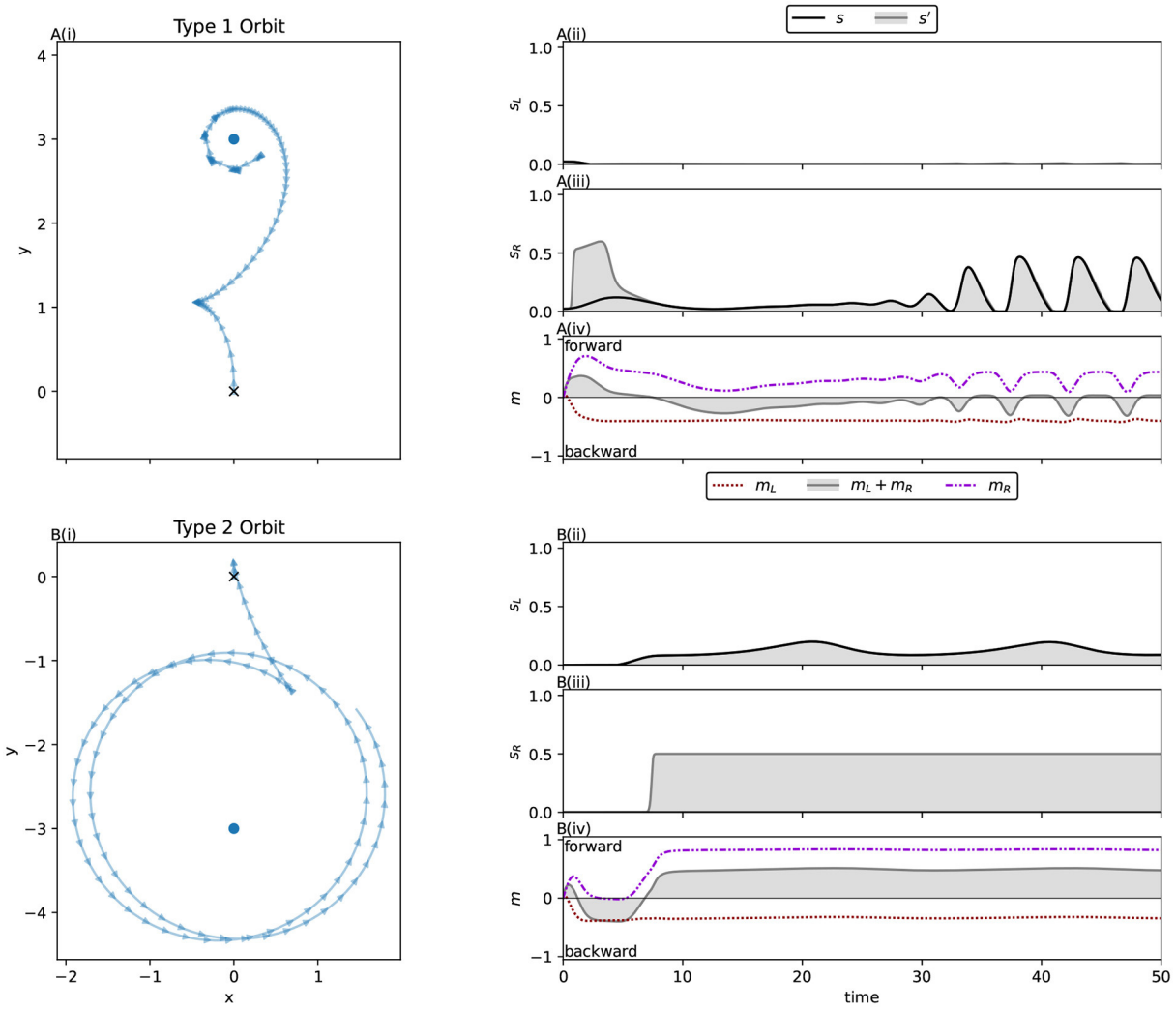



FIGURE 8
Two distinct orbit types produce the bimodal right motor activity distribution seen for the solution evolved with sigmoidal interference in Figure 6B. **(A)** The type 1 orbit, which alternates between driving forwards and backwards to stay close to the light. **(B)** The type 2 orbit, where the robot exclusively drives forwards during the orbit phase. **(i)** The spatial trajectory of the robots, **(ii,iii)** the robots left and right sensor activities respectively, and **(iv)** the robots' left and right motor activations. The black line in **(ii,iii)** shows the environmental stimulation of the sensor, while the grey line and corresponding shaded region shows the total activation of the sensor when both the environmental and motor-driven stimulation are combined. Note the minimization of interference during the Type 1 orbit, in contrast with high level of right sensor interference during the Type 2 orbit.

stimulation is actually constructive in that it synchronizes with and amplifies the environmental stimulus's effect on the sensor. Figure 6B makes it clear that the right motor's activity has not been lowered or even constrained to a tighter range the way the left motor's has—though we still see a slight reduction in interference compared to what the ancestral solution would produce (see Figure 9B). How the system performs so accurately in the presence of this interference becomes clear when we consider the relationship between the right motor activity and the right sensor. As with the ancestor, the robot approaches the light while driving backwards, in such a way that the light enters the right sensor's field from it's blind spot at very close proximity to the sensor. Figure 10A shows an example of this approach. When the light enters the right sensor's field, its activation immediately spikes. In response, the right motor's activity also spikes, causing the robot to drive forwards, and also causing a spike of interference in the same sensor. This is a version of the ancestral Phase B orbit behavior, executed with reduced baseline motor activity, and high right motor activity coordinated with right sensor stimulation. By keeping motor activity at a low baseline and interacting with the environment in such a way that environmental stimuli are sharp and intense, this

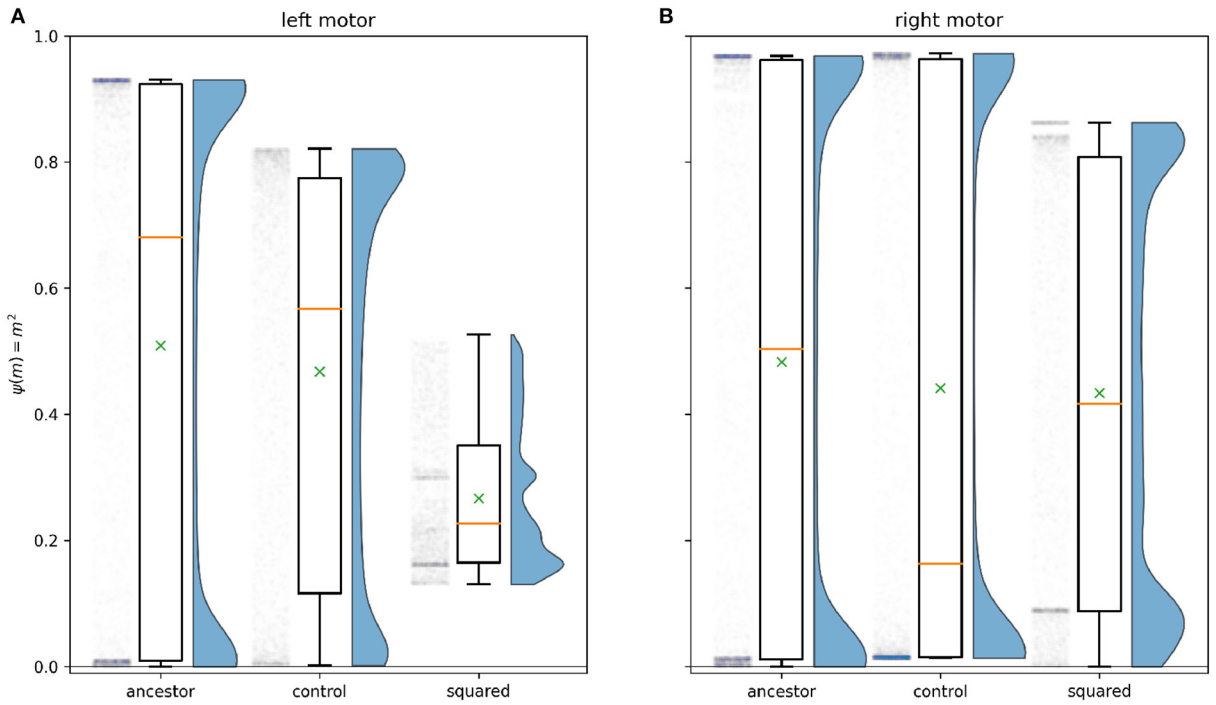



FIGURE 9
Motor-driven interference is reduced in Experiment 2 relative to the ancestral population. The figure shows $\psi(m) = m^2$ for the 12 light coordinates shown in Figure 3, for $20 < t < 50$. **(A)** Note primarily the lowered mean, median and maximum interference with the left motor. Despite the right motor's activity being spread across a wider range than either ancestor or control (see Figure 6), this spread is to low motor activity values, decreasing maximum right motor-sensor interference. **(B)** However, the right motor activity has definitely not been suppressed the way the left has, and the systems successful performance in the presence of this interference ultimately depends on the coordination of right motor-sensor interference with environmental stimulation of the right sensor (see main text).

solution facilitates distinguishing environmental stimuli from low levels of self-caused background noise. By then coordinating motor activity with elevated environmental stimulation of the ipsilateral sensor, motor-driven interference can be raised to high levels without interfering with the system's function, "hiding" in the shadow of the environmental stimulus. Not only does this activity not interfere with perception of the environment, the stimulation caused by right motor's activity actually reinforces and amplifies the environmental stimulus's effect on the sensor above the maximum level it would be able to achieve on its own.

Since right sensor stimulation leads to right motor activity, which in turn leads to more right sensor stimulation, we should address the possibility of a self-sustaining positive feedback loop. This possibility is limited by two forms of negative feedback. The system's relationship to the light source is structured in such a way that elevated right motor activity in response to the environmental stimulus moves the right sensor away from the light, eliminating that stimulus. This is environmentally mediated negative feedback. It is complimented by internal negative feedback. Figure 11A shows how a spike in right sensor stimulation causes an initial strong response in motor activity. However, despite continued stimulation at an elevated level, sufficient to saturate the output of the sensor neuron, motor activity quickly falls from the initial peak. Thus, both internal and environmentally mediated negative feedback play a role in preventing this orbit behavior from being disrupted by motor-driven positive feedback.

As we also saw with sigmoidal interference, this solution realizes a second orbit pattern of Type 2. Positive rather than negative feedback plays a dominant role in this orbit, which comes into effect when the robot is close to the light, but the light is on its left (see Figure 10B). The system's response to left sensor stimulation does not feature the internal negative feedback that right sensor stimulation does, and it produces a response in both right and left motor activity. This in turn produces interference in both sensors. The ultimate effect is that the robot drives forwards in a counter-clockwise orbit around the light. This keeps the left sensor continually stimulated by the light, while the right sensor is continually stimulated by the right motor's activity. In this case we have an environmentally mediated, positive feedback loop, where left sensor stimulation causes the robot to turn toward that stimulus, and the resulting motor-sensor interference produces the same effect.

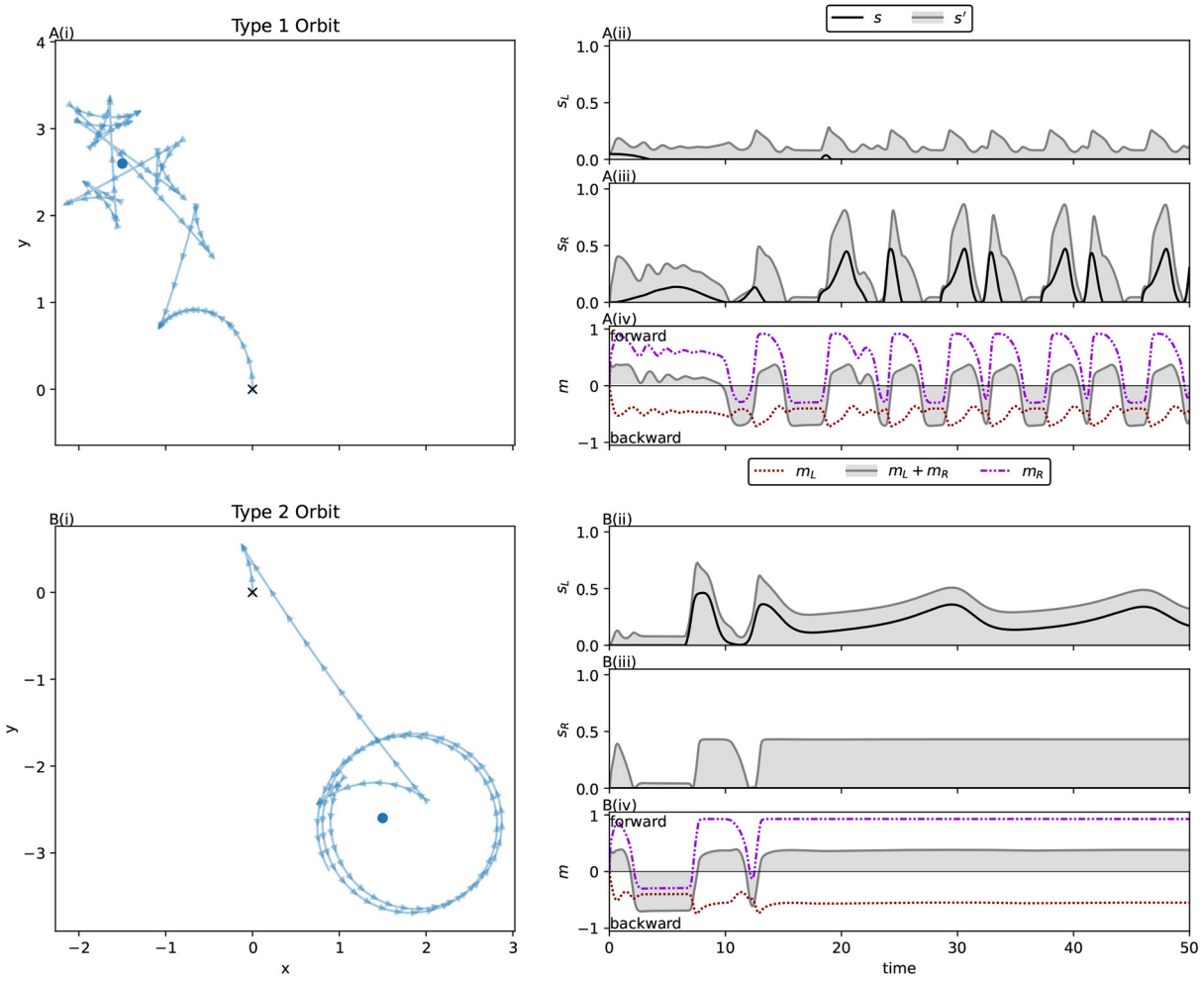



FIGURE 10
Spatial trajectories and sensorimotor activity showing a Type 1 and Type 2 orbit for the solution evolved with squared interference. Subfigures are labeled as in Figure 8. **(A)** Shows a Type 1 orbit reminiscent of the ancestral solution, where motor activity is coordinated with sharp spikes of environmental stimulation of the right sensor. **A(iii)** Shows how elevated right motor interference coincides with environmental right sensor stimulation, amplifying it. The spiking activity is characteristic of negative feedback in this solution, where action resulting from sensor stimulation leads to the stimulus diminishing. **(B)** Shows a Type 2 orbit, where the robot orbits while driving forwards. **B(iv)** Shows how the motor activity plateaus during the orbit, with high right motor interference seen in **B(iii)**. This is associated with positive feedback in this solution, where sensor stimulation leads to activity prolonging that stimulation.

The way this system has been parametrized by the GA relies on the presence of motor-driven stimulation to perform phototaxis. Recall that the ancestor evolved to have zero left sensor activation in many situations, with a left sensor guided approach phase (Phase A) for a number of initial light positions. This trait remains in a way, where the left sensor is often completely free of environmental stimulation, and the left motor activity is constrained to produce lower levels of interference. Nevertheless, this interference plays an important role. Figure 12 illustrates how removing the motor-driven sensor stimulation from just the left sensor causes the approach phase to fail in the majority of cases, succeeding only when its trajectory inadvertently brings it close to the light. This is not unexpected, given that the system was optimized for the presence of motor-driven interference. However, it means that accurate control of the system's motor activity has been optimized in such a way that it now depends on perceiving the direct sensory effects of its own activity. Like the right motor, the left motor responds to sensor stimuli, though in a smaller range and with elevated negative rather than positive activation. This plays an interesting role in the system's response to right sensor stimulation (as in the Type 1 orbit shown in Figure 10A). Note how the coordinated peaks of right environmental and motor-driven sensor stimulation coincide with elevated left motor activity and corresponding motor-driven left sensor stimulation. Figure 11 shows how the presence of left motor-sensor interference amplifies and extends

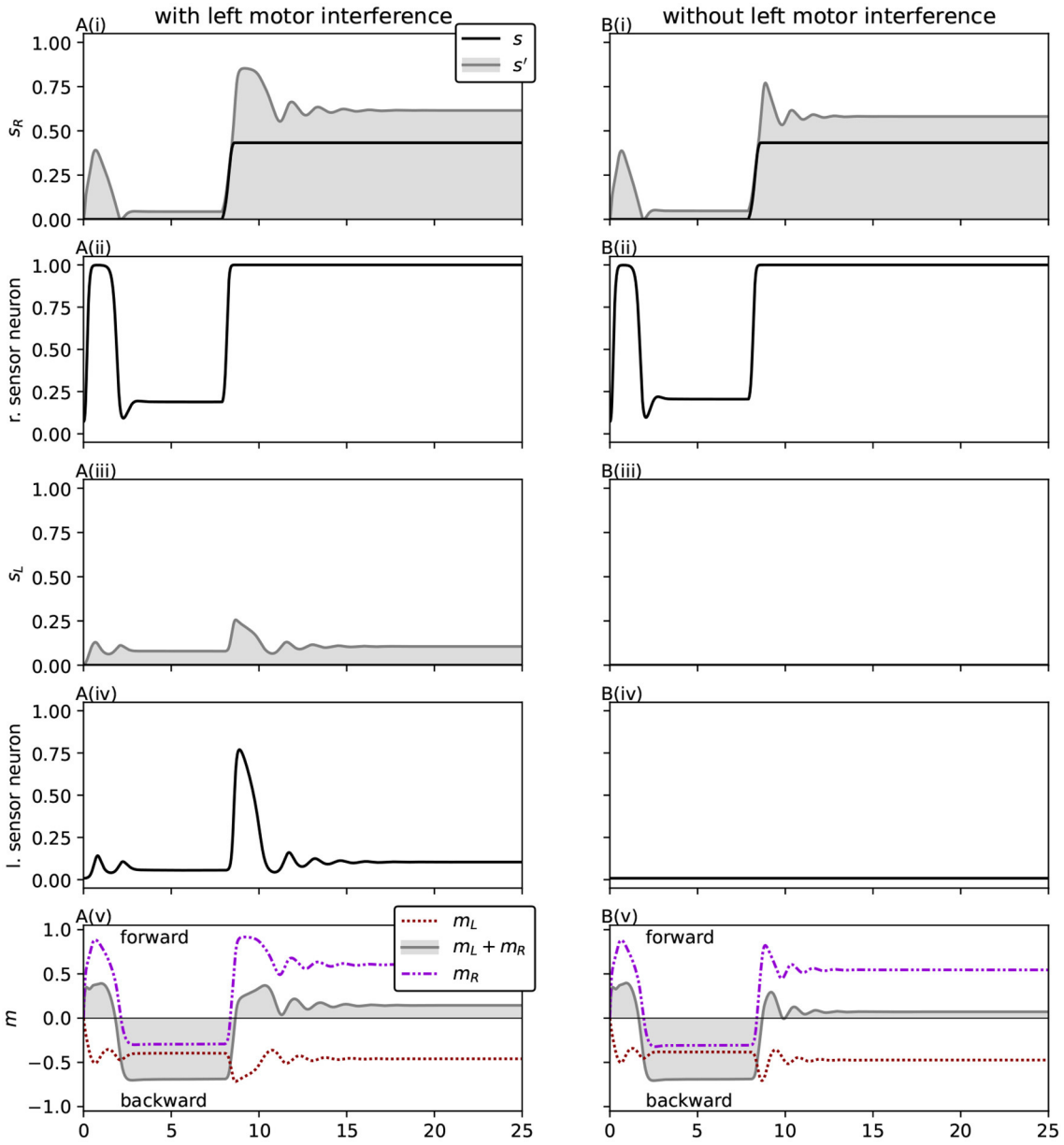



FIGURE 11
The magnitude and duration of the initial motor response to sensor stimuli are strengthened by the presence of left motor interference. Sensorimotor activity and sensory neuron output time series are shown for the solution evolved with squared interference (Equation 13), when the right sensor is presented with an artificial environmental stimulus, which spikes and plateaus around $t = 8$. **(A)** Shows the response under the condition of evolutionary adaptation for the robot, with motor interference present. **(B)** Shows the response when the left motor-sensor interference is removed. The duration and intensity of the motor response to the stimulus is diminished without the interference, indicating that the interference plays a functional role in the evolved behavior. Additionally, it can be seen that the response to sudden right sensor stimulation is accompanied by internal negative feedback—even when the stimulation persists, motor activity quickly falls from the initial peak.

the initial motor activity response to right sensor stimulation. This demonstrates not only a specific way in which the system has been optimized for the presence of interference, but also how self-caused stimuli can play a directly functional role in behavior.

To summarize, we see the ancestral strategy is well preserved in this evolved solution. This solution can be characterized as minimizing interference to an extent, as we also saw in the case of sigmoidal interference. We also see a condition where motor-driven sensor interference does not need to be minimized, namely when it can be made to coincide temporally with environmental stimulation of the same sensor. Here the onset of the environmental stimulus prompts the

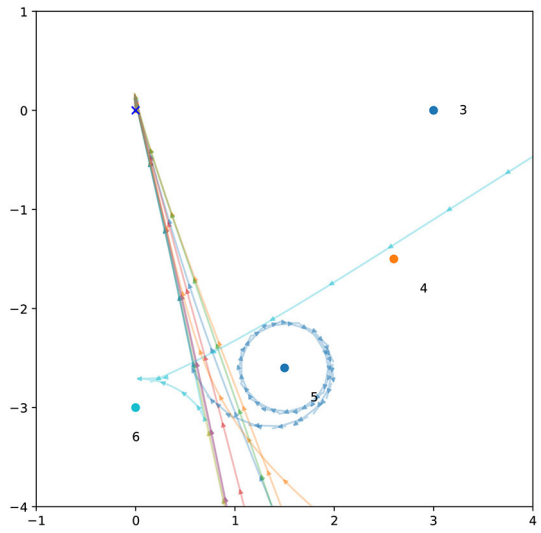



FIGURE 12
When motor-driven interference is removed, the behavior evolved with squared interference fails. Spatial trajectories for 12 light coordinates (Figure 3) are plotted with all motor-sensor interference removed. The approach phase now only succeeds in two out of 12 cases, where the blind approach brings the robot close to the light. The orbit phase only succeeds in one of these two cases.

interfering motor activity, and a combination of internal and environmentally mediated negative feedback extinguishes both interfering activity and stimulus. In this case the motor-driven stimulation does not interfere with perception of the environmental stimulus, instead reinforcing and amplifying it. This obviates the need to distinguish or subtract the self-caused stimulus from the environmental. Separately, we also see that a stable, periodic orbit phase can be facilitated by positive feedback. Finally, we found that while left motor-sensor interference is confined to a narrow range, the system has been optimized to rely on its presence and even incorporate it functionally.

## 3.4. Experiment 4: Time dependent interference

With both of the preceding interference functions, if the motor activity is held constant, then the interference will also take on a constant value. Since the interference is additive and non-saturating, subtracting a constant term can remove the interference and leave only the environmental signal—no prediction required. In general a CTRNN with a sufficiently high bias $\beta$ for the input neurons can do this, though in our case the maximum value we permit the GA to assign to $\beta$ is too low to fully compensate for maximal interference. Nevertheless, solutions to the previous two interference functions have shown both the utility of avoiding or minimizing motor-sensor interference, as well as the role that holding motor activity and its corresponding interference constant can have in constructing long-term stable relationships with environmental sources of sensor stimulation. With the following function it is not possible for the interference to plateau at a constant value. It describes a sine wave with a maximum of 1 and a minimum of 0, whose frequency is determined by the motor activation:

$$\psi = \frac{\sin(c) + 1}{2} \tag{14}$$

$$\dot{c} = (b + |m|)r \tag{15}$$

Here $c$ gives the phase of the sinusoidal, capturing the previous values of $m$. $b = 0.1$ determines the base frequency of the sinusoidal in the absence of any motor activity, while $r = 8$ is the frequency range term determining the maximum frequency the sinusoidal can reach. The effect of adding 1 and dividing by 2 is simply to shift the wave from the range $[-1, 1]$ to the range $[0, 1]$. This equation essentially advances through a standard sine wave at a rate determined by the motor activity. As with the previous interference functions, the interference for a given sensor is calculated from the ipsilateral motor, such that when computing the interference for the left sensor we have $m = m_L$, and for the right sensor $m = m_R$.

Unlike the previous interference functions, this is not purely a function of the motor activity, such that if you know $m$ at time $t$, you know $\psi$ at time $t$. Instead it is a function of time, depending on the prior history of the system, specifically on all the previous motor activity up to the current time. More importantly for our purposes, if the input is held constant, the output continues to vary over time. We will refer to the interference generated by Equation (14) as *time dependent* or *sinusoidal* interference. A trial duration of 20 time units was used during evolution for this interference function.

Using this time dependent interference function we find that while avoiding interference, minimizing it, or holding it constant are all important ways of coping with self-caused stimuli, they are not the only ways. Timescale differences between the frequency of the motor-driven interference and the frequency of environmental stimulation of the sensor can be exploited to distinguish the two, and behavior can shape both interference and environmental stimuli to amplify these differences.

In this system the environmental signal is able to be detected despite the presence of interference, due to differences in timescale between the motor-driven interference and the frequency of environmental stimulation of the sensors. First let's demonstrate that the system actually can respond to environmental stimuli. Figure 13 illustrates how a spike in environmental stimulation of the left sensor has an excitatory effect on both motors, causing the system to switch from driving backwards to driving forwards. Observing the behavior of the

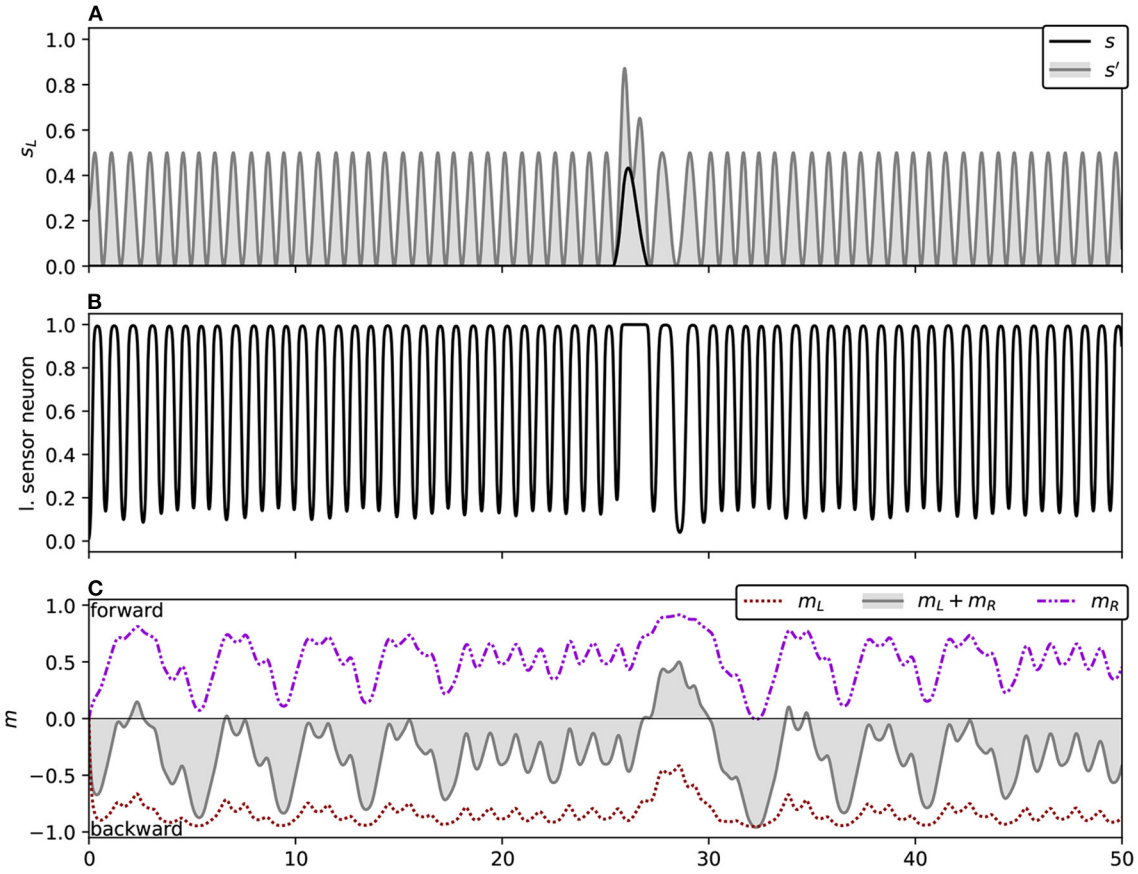



FIGURE 13
Sensorimotor activity and sensory neuron output time series are shown for the solution evolved with sinusoidal interference (Equation 14), **(A)** The left sensor is presented with a spike in environmental stimulation at around $t = 28$. **(B)** The neural response to the environmental stimulus is clearly visible—prolonged saturation of the left sensor neuron's output function (see Equation 5). **(C)** The spike of environmental sensor stimulation causes the robot to drive forward instead of backwards for a time, demonstrating that the system can respond to environmental stimuli.

output functions of this system's two sensor neurons, we found elevated neural biases $\beta$ compared to the ancestral solution: remembering that $-5 \leq \beta \leq 5$, we observe 4.67 and 3.73 for the left and right motor, respectively, compared to –0.75 and 0.99 in the ancestral solution. These sensor neuron biases are calibrated such that (A) with no environmental stimulation, the neuron's output function is maximized only with the peaks of the sinusoidal interference, and (B) when combined with sufficient environmental stimulation, the troughs of the sinusoidal interference are high enough that the output function is maximized continually. This can be seen in the neural response to environmental stimulation shown in Figure 13B. This makes the environmental signal detectable despite the continuously varying interference. This solution is made possible by the large difference in timescale between the frequency of the sinusoidal interference and the frequency with which the sensor receives the environmental stimulation. In this system, the frequency of the interference can be an order of magnitude higher than the frequency of environmental stimulation, as can be seen in Figure 14. This difference in timescale means that the minimum value of the sinusoidal interference is bound to coincide multiple times with each period where there is no environmental sensor stimulation. This means that a drop in neuron firing always coincides with the absence of environmental sensor stimulation, so over time the system can reliably respond to environmental stimuli.

While the evolution of our model was constrained in such a way that it could not implement it, there is another solution for filtering out interference of a sufficiently high timescale relative to the frequency of environmental sensor stimulation that peak interference is guaranteed to coincide with all instances of environmental stimulation. The maximum bias of nodes in our model was constrained to the maximum weight of a single incoming connection (5), which is lower than the product of the environmental intensity factor with the input scaling factor applied to inputs to the sensor neurons ($5 \times 5 = 25$). However,

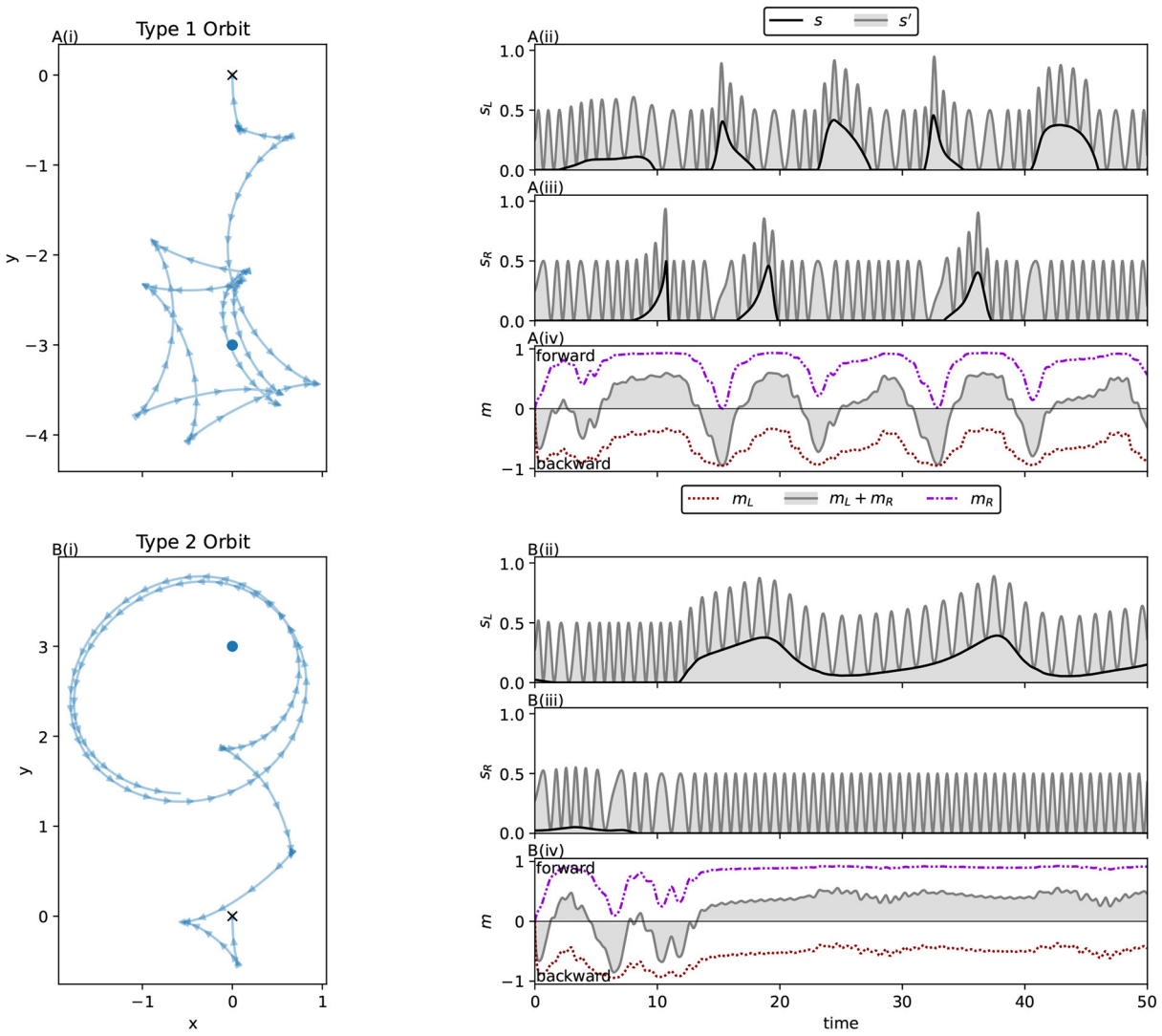



FIGURE 14
Spatial trajectories and sensorimotor activity for the solution evolved with squared interference. Subfigures are labeled as in Figure 8. The sensor plots show how the relatively slowly changing environmental sensor stimulation raises the minima of the high frequency interference, allowing the environmental stimulus to be responded to despite the interference. The difference in timescale that makes this possible is clearly visible here. Responsiveness to the environment is most clearly visible in **A(iv)**, where more positive motor activity is associated with environmental stimulation of the left or right sensor. The continual oscillations in motor activity (most clearly visible in the gray net motor activity line) are driven by the high frequency interference. These oscillations produce the elliptical Type 2 orbit seen in **B(i)**.

a sufficiently high bias (around 12) can indeed induce the sensor neurons' output function to only be maximized when environmental stimulation is high.

These two ways of adjusting the neural biases demonstrate how a large difference in timescale between environmental signal and interference means that over time it is possible to extract the environmental signal from the summation of the two. However, such differences in timescale are not guaranteed, and it is here that the embodied nature of this system comes into play. The robot's motor activity actually amplifies any pre-existing difference in timescale, as typical motor activity is constrained to higher absolute ranges than the ancestral solution—see Figure 6.

Due to the way this time dependent interference periodically saturates the input neurons, the system is not sensitive to environmental stimuli spikes that are of sufficiently low duration to perfectly coincide with motor interference peaks as the corresponding input neuron's output function would already be saturated. Note that spikes of this duration do reliably induce a motor response in the other systems we've examined in this paper. This represents a problem for the ancestral solution's strategy of taking advantage of sharp spikes in the right sensor. Significantly—and despite the system's elevated right motor activity—this system's Type 1 orbit is much slower than the ancestor's, with the periods of environmental stimulation of

the sensor lasting for longer. This avoids the problem of the environmental stimulus being too short duration, and further amplifies the differences in time scale. So when it comes to distinguishing environmental and self-caused stimuli, the motor activity of the system not only shapes the self-caused stimuli to facilitate this, it shapes the environmental stimuli too.

As with the unavoidable squared interference, the behavior of this system depends on the presence of its motor-driven interference. For example, with the left motor-sensor interference removed, environmental stimulation of the left sensor inhibits rather than excites the activation of both motors. Significantly, in the absence of environmental stimulation, the motor activity and corresponding interference of this system features a long transient before settling into lower magnitude oscillations, and this transient is restarted by environmental sensor stimulation. This effect can be seen in Figure 13. These prolonged effects of momentary environmental stimulation are not seen in the systems examined in Experiments 1–3. They mean that the frequency of the motor-driven interference varies significantly both during the approach to the light and during Type 1 orbits. Altogether these qualities demonstrate that the evolved behavior of this system depends on its motor-driven interference, emphasizing that even interference as seemingly unruly as this can be incorporated into successful behavior.

To summarize, this system has the ability to respond to environmental stimuli despite continually varying sinusoidal interference. Rather than subtracting out the motor-driven interference, the behavior of the system is deeply entangled with it, displaying oscillatory motor activity driven by the interference and prolonged transient motor activity following activation of the motors in response to stimuli. Additionally, whether an environmental stimulus is excitatory or inhibitory depends, respectively on the presence or absence of motor-driven sensor stimulation. This demonstrates that rather than suppressing self-caused stimuli, proper functioning for some systems relies on the presence of self-caused stimuli. In this system we see responsiveness to the environment facilitated by a fixed solution that is implemented at the evolutionary timescale, rather than prediction and subtraction of self-caused stimuli on the timescale of actions. Because of the difference in timescale between the frequency of the sinusoidal interference and the frequency of environmental stimulation, a CTRNN neuron can be parametrized such that the maximization of its output function only coincides with environmental sensor stimulation, or such that the minimization of its output function only coincides with the absence of such stimulation. Most significantly for the role of embodiment in coping with self-caused sensory stimuli, we see that this difference in time scale between motor-driven and environmental sensor stimulation is amplified by the system's behavior, which both elevates the frequency of motor-driven sensory stimulation and lowers the frequency of environmental sensor stimulation.

# 4. Discussion

One explanation of the sensory attenuation effect is that self-caused sensory stimuli are predicted internally using a copy of the relevant neural outputs, and then subtracted out of the sensory inputs (Wolpert et al., 1995; Miall and Wolpert, 1996; Roussel et al., 2013; Klaffehn et al., 2019). This may well be the case, but even in a model where this predict-and-subtract mechanism would be a perfect solution, our GA instead found other viable alternatives. We have shown that a neural network controller can be successfully adapted to handle several different forms of motor-driven sensory interference, and significantly, the adaptations we have cataloged here do not rely on *predicting* this interference. We now summarize these adaptations.

**Avoidance:** When self-caused sensory interference is only triggered by certain motor outputs, and if the task at hand can be accomplished while avoiding those outputs, it may be easiest for a control system to simply modify its behavior to avoid motor-sensor interference. We saw this emerge when our model was evolved with sigmoidal interference. It is not clear whether we should expect this avoidance approach to scale well to a more numerous and complex arrangement of sensors and motors, though it seems that the problem of prediction would also become more complex in such circumstances. In the special case where there are multiple independent sensors and motors, where each motor interferes with only one sensor, an alternative solution is possible. If the task can be accomplished using only one sensor, then only one source of interference needs to be regulated. Doing so permits the other motors to operate freely over a wider range of activity. We describe this strategy as "keeping one eye on the prize". This is arguably just avoiding the interference, with extra steps. We again saw this strategy used in the case of sigmoidal interference.

Where interference is unavoidable but the magnitude of the interference does depend on motor activity, motor activity can be constrained to ranges that limit the quantity of interference, reducing its magnitude relative to environmental stimuli. This is used in the case of the unavoidable squared interference.

Minimization and avoidance could be seen as special cases of causing the interference to plateau at a constant value. If interference is additive and non-saturating, as it is in our model, it can be eliminated by simply subtracting a constant term from the input. In general this is trivial for a CTRNN. However even without subtracting the interference out directly, constant interference just shifts an environmental stimulus's contribution to the sensor to a higher range, which does not actually change the information available when the interference is non-saturating.

**Coordination:** The timing of motor-driven interference with a sensor may be regulated to coincide with environmental stimulation of that same sensor. One way to look at this

is that the detection of a sufficiently 'loud' environmental stimulus renders any coincident interference irrelevant. With a one dimensional sensor like those used in this model, the interference is actually constructive, that is the coincidence of motor-driven and environmental stimuli amplifies the effect of the environmental stimulus on the sensor. If the response to such a stimulus tends to diminish that stimulus (negative feedback), as we see when stimulation of the sensor causes the robot to turn away from the light, then this strategy of coordination can play a powerful role in establishing a stable relationship with environmental stimuli. This can be effectively combined with a strategy of avoiding or minimizing interference, which we saw with the squared interference function. The constructive interference we saw here may not be possible with more complex collections of sensors, where environmental and self-caused stimulation do not interact as straightforwardly as in our model. This is not to say that non-predictive, embodied solutions would not be found in such situations. On the contrary, discovering solutions afforded by richer embodiments may be a fruitful avenue for future work.

**Time scale differences:** The previous solutions don't work for interference which is continually varying in such a manner that the interference's minima and maxima are not under direct control of the motors. However, if such interference is of a high enough frequency relative to the frequency of environmental sensor stimulation, then this difference in time scale can be leveraged to separate interference from environmental stimuli. Slowly varying stimuli can be perceived through quickly varying interference, which we saw with the sinusoidal interference function. The evolved behavior we saw with this interference function elevated the frequency of motor-driven stimulation further, amplifying this differential.

**Shaping environmental stimuli:** Time scale differences are a case of natural differences between the characteristics of the interference and the environmental stimuli. So far we've described how the system can shape the interference to minimize its negative effects or make it easier to distinguish from the environmental stimuli. However, the ancestral solution demonstrates that the shape that environmental sensor stimulation takes depends on the system's activity—sharp spikes in sensor stimulation are produced by passing close to the light while driving backwards. With the sinusoidal interference function, we found that sharp spikes could be lost in the high frequency interference, and that in addition to the system's behavior raising the frequency of the motor-driven interference, its behavior also lowered the frequency of environmental stimulation. Embodied systems can reliably respond differently to environmentally and self-caused stimuli because the characteristics of both forms of stimuli are at least partially determined by the system's own activity.

Removing motor-driven interference from a system optimized to perform a task in the presence of that interference does not necessarily improve performance, and may instead degrade it significantly. Instead the successful phototactic behavior of the systems we've studied often incorporates interference functionally. Coordination of interference with environmental sensor stimulation is one case of this, where the coordination amplifies the stimulus, but we also saw how the response to environmental stimulation of one sensor can be mediated by motor-driven stimulation of the contralateral sensor. This suggests that it is a mistake to view the problem of coping with self-caused sensory stimuli as primarily about subtracting out the interference—even viewing it in terms of perceiving the environment clearly despite the interference may be going too far. It's natural to think of the phototaxis task this way, but the evolutionary algorithm we used selected purely for phototactic ability, and as we've seen this can involve incorporating motor-driven interference into behavior. Despite our attempt to set up a model and problem where sensory attenuation is a perfect solution, the solutions cataloged here for coping with self-caused sensory interference do not align with the sensory attenuation phenomena that has been studied experimentally (e.g., Pareés et al., 2014), raising the broad question of what conditions would lead to sensory attenuation emerging.

This all reinforces that prediction and subtraction cannot tell the whole story when it comes to coping with self-caused sensory stimuli. In some ways this is obvious, as self-caused sensory stimuli are involved in a range of activities in which they do not play an interfering role. For example, the sensation of self-touch when kneading an aching muscle, or occlusion of the visual field when engaging in visually guided reaching and grasping. In these activities, self-caused sensory stimuli are actually desirable. Nevertheless, our model shows that even in situations where clear perception of the environment is *prima facie* desirable, self-caused sensory stimuli may not play an entirely interfering role. Furthermore, we see that even when responsiveness to the environment is needed, prediction and subtraction are not the only game in town.

How do these results actually relate to the predictive account of coping with self-caused stimuli? A criticism of our results may be that the problems being solved in our model are insufficiently "representation-hungry" to require prediction. Representation-hungry problems are those that seem to require the use of internal representations to be solved, defined by Clark and Toribio (1994) to be a problem where one or both of the following conditions hold. Condition one is that the problem involves reasoning about *absent*, non-existent, or counterfactual states of affairs. Condition two is that the problem demands selective sensitivity to parameters whose sensory manifestations are "complex and unruly" - that is, the system must be able to treat differently inputs whose sensory manifestations are highly similar, and conversely be able to treat similarly inputs whose sensory manifestations are very different. We actually agree that our model does not solve a representation-hungry problem, and in fact see this is a primary contribution of our results. In

general, coping with self-caused sensory stimuli need not be a representation-hungry problem.

How we process self-caused stimuli is often taken to involve an internal predictive model (e.g., Roussel et al., 2013; Klaffehn et al., 2019). Prediction itself is a task which meets Clark and Toribo's first condition, since prediction inherently involves states of affairs that do not yet exist. However, the fundamental problem the predictive model is being used to solve meets only the second criteria, that is treating differently self-caused and externally-caused inputs whose sensory manifestations may be identical. Otherwise identical inputs can be distinguished by predicting, based on an internal model, whether an input is self-caused or externally-caused. If prediction is necessary then the problem of coping with self-caused stimuli would seem to meet the criteria for representation-hunger.

The way self-caused stimuli have been studied experimentally highlights what we see as a key limitation of the representational paradigm. Experiments such as force matching (intentionally and justifiably) aim to isolate specific psychological phenomena and neural mechanisms. We would like to suggest that doing so may naturally lead to overemphasizing the role of these studied mechanisms when extrapolating explanations back from the experiment to real world behavior. Specifically, a limitation of the force matching experiment is the highly constrained motor outputs of the subject—the subject is responding to one specific stimulus (force applied to a finger) with a very limited range of motor outputs, either pressing on that finger or moving a mechanism with their other hand (Pareés et al., 2014). In contrast, coping with analogous perceptual problems in the "real world" might tend to take advantage of their less constrained sensori*motor* coupling with the environment—but this wouldn't show up in force matching experiments. This is not a criticism of the experiments, but we do suggest that evolutionary robotics models like this one can help highlight that behaviors depending on a more dynamical, ongoing, and open ended context may play important roles in problem-solving, which may not manifest clearly in the deliberately restricted range of sensorimotor interactions possible in tightly controlled experiments. Under laboratory conditions, a strict interpretation of Clark and Toribio's second criteria may hold—where self-caused and externally-caused stimuli are identical to the extent that only knowledge over and above their sensory manifestations can distinguish them. However, the everyday problem of coping with self-caused sensory stimuli occurs outside the lab, where these stimuli are part of our ongoing sensorimotor activity. In this case our model has shown that there are diverse ways to perform successfully and even to disentangle self-caused and externally-caused stimuli. A key part of this is that both types of stimuli are shaped by our own activity, and thus encountered on our own terms. In these circumstances, the strict definition is unlikely to hold, as we can shape both self and externally caused stimuli to differentiate them.

While the problem of distinguishing truly identical sensory inputs may well be representation-hungry, our model's embodiment allows it to shape its inputs such that they are distinguishable by non-predictive means. Thus, we grant that our model does not capture a strictly representation-hungry problem, a conclusion directly supported by our results. This is a not a limitation of this study, it's a feature. Our model shows that representational cognition is not necessary in general to cope with self-caused stimuli, because of the capabilities afforded by embodiment. In effect, this shrinks the set of human capabilities which are taken to require representational cognition.

The idea of representation-hunger highlights a long running critique of embodied cognition, where solving tasks in representation-free, embodied ways aren't considered central examples of what we really mean by cognition. A distinction is drawn between tasks solvable *via* online and potentially representation-free sensorimotor processing, and offline cognition operating on internal, representational models (Zahnoun, 2019). It is worth noting that similarly minimal, CTRNN controlled models have successfully solved problems with requirements like memory without the use of internal representations. Beer and Williams (2015) demonstrate how a robot can both remember a cue and categorize a subsequent probe relative to that cue by offloading memory to the environment and structuring its relationship with its environment to facilitate direct perception on the relative difference between cue and probe. It was only when the robot's ability to move while being presented with the cue was removed that information about the cue was retained internally in the neural activation. Studies like this push back at the idea that internal representation is necessary to solve problems requiring responses to abstract or absent stimuli, by showing that other possibilities are facilitated by the way embodiment structures the ongoing relationship between controller and environment.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

JG and ME conceived of the idea for this project together. JG wrote the code, ran the experiments, analyzed the data, produced the figures, and wrote the manuscript. ME provided feedback on the experiments and on multiple drafts of the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This research was funded by a University of Auckland Doctoral Scholarship. OAP fees were funded by the School of Computer Science, University of Auckland.

## Acknowledgments

The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and/or training services as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL https://www.nesi.org.nz. Figures were produced with Matplotlib (Hunter, 2007). The idea for the rain-cloud presentation of the data used in Figure 6 is due to Allen et al. (2019). An extended abstract summarizing parts of this work was presented at the ALIFE 2022 conference (Garner and Egbert, 2022).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., and Kievit, R. A. (2019). Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res*. 4, 63. doi: 10.12688/wellcomeopenres.15191.1

Bays, P. M., Wolpert, D. M., and Flanagan, J. R. (2005). Perception of the consequences of self-action is temporally tuned and event driven. *Curr. Biol.* 15, 1125–1128. doi: 10.1016/j.cub.2005.05.023

Beer, R. D. (1996). "Toward the evolution of dynamical neural networks for minimally cognitive behavior," in *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, eds P. Maes, M. Mataric, J. Meyer, J. Pollack, and S. Wilson (MIT Press), 421–429.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adapt Behav*. 11, 209–243. doi: 10.1177/1059712303114001

Beer, R. D. (2006). Parameter space structure of continuous-time recurrent neural networks. *Neural Comput*. 18, 3009–3051. doi: 10.1162/neco.2006.18.12.3009

Beer, R. D., and Williams, P. L. (2015). Information processing and dynamics in minimally cognitive agents. *Cogn. Sci*. 39, 1–38. doi: 10.1111/cogs.12142

Blakemore, S.-J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *Neuroreport* 11, R11-R16. doi: 10.1097/00001756-200008030-00002

Blakemore, S.-J., Wolpert, D. M., and Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nat. Neurosci*. 1, 635–640. doi: 10.1038/2870

Chatila, R., Renaudo, E., Andries, M., Chavez-Garcia, R.-O., Luce-Vayrac, P., Gottstein, R., et al. (2018). Toward self-aware robots. *Front. Rob. AI* 5, 88. doi: 10.3389/frobt.2018.00088

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci*. 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, A., and Toribio, J. (1994). Doing without representing? *Synthese* 101, 401–431. doi: 10.1007/BF01063896

Crapse, T. B., and Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nat. Rev. Neurosci*. 9, 587–600. doi: 10.1038/nrn2457

de Lange, F. P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception? *Trends Cogn. Sci*. 22, 764–779. doi: 10.1016/j.tics.2018.06.002

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci*. 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Garner, J., and Egbert, M. (2022). "Is prediction required? using evolutionary robotics to investigate how systems cope with self-caused sensory stimuli," in *ALIFE 2022: The 2022 Conference on Artificial Life* (MIT Press).

Harvey, I. (2011). "The microbial genetic algorithm," in *Advances in Artificial Life. Darwin Meets von Neumann. ECAL 2009. Lecture Notes in Computer Science*, Vol. 5778, eds G. Kampis, I. Karsai, and E. Szathmáry (Berlin; Heidelberg: Springer). doi: 10.1007/978-3-642-21314-4_16

Harvey, I., Paolo, E. D., Wood, R., Quinn, M., and Tuci, E. (2005). Evolutionary robotics: a new scientific tool for studying cognition. *Artif. Life* 11, 79–98. doi: 10.1162/1064546053278991

Hughes, G., and Waszak, F. (2011). ERP correlates of action effect prediction and visual sensory attenuation in voluntary action. *Neuroimage* 56, 1632–1640. doi: 10.1016/j.neuroimage.2011.02.057

Hunter, J. D. (2007). Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi: 10.1109/MCSE.2007.55

Kahl, S., Wiese, S., Russwinkel, N., and Kopp, S. (2022). Towards autonomous artificial agents with an active self: modeling sense of control in situated action. *Cogn. Syst. Res*. 72, 50–62. doi: 10.1016/j.cogsys.2021.11.005

Kilteni, K., and Ehrsson, H. H. (2017). Body ownership determines the attenuation of self-generated tactile sensations. *Proc. Natl. Acad. Sci. U.S.A*. 114, 8426–8431. doi: 10.1073/pnas.1703347114

Kilteni, K., and Ehrsson, H. H. (2022). Predictive attenuation of touch and tactile gating are distinct perceptual phenomena. *iScience* 25, 104077. doi: 10.1016/j.isci.2022.104077

Kilteni, K., Engeler, P., and Ehrsson, H. H. (2020). Efference copy is necessary for the attenuation of self-generated touch. *iScience* 23, 100843. doi: 10.1016/j.isci.2020.100843

Klaffehn, A. L., Baess, P., Kunde, W., and Pfister, R. (2019). Sensory attenuation prevails when controlling for temporal predictability of self- and externally generated tones. *Neuropsychologia* 132, 107145. doi: 10.1016/j.neuropsychologia.2019.107145

Lalouni, M., Fust, J., Vadenmark-Lundqvist, V., Ehrsson, H. H., Kilteni, K., and Birgitta Jensen, K. (2021). Predicting pain: differential pain thresholds during self-induced, externally induced, and imagined self-induced pressure pain. *Pain* 162, 1539–1544. doi: 10.1097/j.pain.0000000000002151

Mathayomchan, B., and Beer, R. D. (2002). Center-crossing recurrent neural networks for the evolution of rhythmic behavior. *Neural Comput*. 14, 2043–2051. doi: 10.1162/089976602320263999

Miall, R. C., and Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Netw*. 9, 1265–1279. doi: 10.1016/S0893-6080(96)00035-4

Pareés, I., Brown, H., Nuruki, A., Adams, R. A., Davare, M., Bhatia, K. P., et al. (2014). Loss of sensory attenuation in patients with functional (psychogenic) movement disorders. *Brain* 137, 2916–2921. doi: 10.1093/brain/awu237

Phattanasri, P., Chiel, H. J., and Beer, R. D. (2007). The dynamics of associative learning in evolved model circuits. *Adapt Behav*. 15, 377–396. doi: 10.1177/1059712307084688

Press, C., Kok, P., and Yon, D. (2020). The perceptual prediction paradox. *Trends Cogn. Sci*. 24, 13–24. doi: 10.1016/j.tics.2019.11.003

Roussel, C., Hughes, G., and Waszak, F. (2013). A preactivation account of sensory attenuation. *Neuropsychologia* 51, 922–929. doi: 10.1016/j.neuropsychologia.2013.02.005

Schillaci, G., Ritter, C.-N., Hafner, V. V., and Lara, B. (2016). Body "Representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents," in *Proceedings of the Artificial Life Conference 2016* (Cancun: MIT Press), 390–397.

Thompson, A., Layzell, P., and Zebulum, R. (1999). Explorations in design space: unconventional electronics design through artificial evolution. *IEEE Trans. Evolut. Comput*. 3, 167–196. doi: 10.1109/4235.788489

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Curr. Biol*. 11, R729-R732. doi: 10.1016/S0960-9822(01)00432-8

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269, 1880–1882. doi: 10.1126/science.7569931

Zahnoun, F. (2019). On representation hungry cognition (and why we should stop feeding it). *Synthese* 198, 267–284. doi: 10.1007/s11229-019-02277-8

frontiers | Frontiers in Neurorobotics

TYPE Hypothesis and Theory
PUBLISHED 15 December 2022
DOI 10.3389/fnbot.2022.880724

Check for updates

OPEN ACCESS

EDITED BY
Inês Hipólito,
Humboldt University of Berlin, Germany

REVIEWED BY
Chris Fields,
Independent Researcher, Caunes Minervois, France
Adam Safron,
Johns Hopkins Medicine, United States

*CORRESPONDENCE
David Harrison
djh253@cam.ac.uk
Wiktor Rorot
w.rorot@uw.edu.pl

RECEIVED 21 February 2022
ACCEPTED 31 October 2022
PUBLISHED 15 December 2022

CITATION
Harrison D, Rorot W and Laukaityte U (2022) Mind the matter: Active matter, soft robotics, and the making of bio-inspired artificial intelligence. *Front. Neurorobot.* 16:880724. doi: 10.3389/fnbot.2022.880724

COPYRIGHT
© 2022 Harrison, Rorot and Laukaityte. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Mind the matter: Active matter, soft robotics, and the making of bio-inspired artificial intelligence

David Harrison[1,2,3]*, Wiktor Rorot [4]* and Urte Laukaityte[5]

[1]Department of History and Philosophy of Science, University of Cambridge, Cambridge, United Kingdom, [2]Leverhulme Centre for the Future of Intelligence, Cambridge, United Kingdom, [3]Konrad Lorenz Institute for Evolution and Cognition Research, Vienna, Austria, [4]Human Interactivity and Language Lab, Faculty of Psychology, University of Warsaw, Warsaw, Poland, [5]Department of Philosophy, University of California, Berkeley, Berkeley, CA, United States

Philosophical and theoretical debates on the multiple realisability of the cognitive have historically influenced discussions of the possible systems capable of instantiating complex functions like memory, learning, goal-directedness, and decision-making. These debates have had the corollary of undermining, if not altogether neglecting, the materiality and corporeality of cognition—treating material, living processes as "hardware" problems that can be abstracted out and, in principle, implemented in a variety of materials—in particular on digital computers and in the form of state-of-the-art neural networks. In sum, the matter *in se* has been taken not to matter for cognition. However, in this paper, we argue that the materiality of cognition—and the living, self-organizing processes that it enables—requires a more detailed assessment when understanding the nature of cognition and recreating it in the field of embodied robotics. Or, in slogan form, that the matter matters for cognitive form and function. We pull from the fields of Active Matter Physics, Soft Robotics, and Basal Cognition literature to suggest that the imbrication between material and cognitive processes is closer than standard accounts of multiple realisability suggest. In light of this, we propose upgrading the notion of multiple realisability from the standard version—what we call 1.0—to a more nuanced conception 2.0 to better reflect the recent empirical advancements, while at the same time averting many of the problems that have been raised for it. These fields are actively reshaping the terrain in which we understand materiality and how it enables, mediates, and constrains cognition. We propose that taking the *materiality* of our embodied, precarious nature seriously furnishes an important research avenue for the development of embodied robots that autonomously value, engage, and interact with the environment in a goal-directed manner, in response to existential needs of survival, persistence, and, ultimately, reproduction. Thus, we argue that by placing further emphasis on the soft, active, and plastic nature of the materials that constitute cognitive embodiment, we can move further in the direction of autonomous embodied robots and Artificial Intelligence.

KEYWORDS

**multiple realisability, fine-grained functionalism, functionalism, soft robotics, active matter physics, basal cognition, artificial intelligence, embodied cognition**

## Introduction

Standard approaches to understanding cognition—and the wider goal of recapitulating it on simulated platforms or in the field of robotics—have tended to neglect the importance of the materiality of the body and its relevance for constraining, enabling, and mediating cognition. This contention is centred right at the origin of the cognitive sciences and is typically framed in terms of *multiple realisability*. It is often argued, then, that cognition is a species of software that, in principle, is instantiable in any back-of-the-envelope set of materials so long as they are "suitably organized" (Putnam, 1975). As Hilary Putnam once put it, "We could be made of Swiss cheese and it wouldn't matter" (Putnam, 1975: 291). Although few authors would defend this version of multiple realisability (MR) today (see Polger and Shapiro, 2016 for a state-of-the-art discussion of the philosophical literature), the belief that the materiality of cognition is mostly a "hardware" problem with the truly interesting explanandum being cognitive "software" that sits above still permeates much of the theoretical and philosophical literature. However, as we will see below, much turns on what it means to be "suitably organized" and it is by no means clear that any pell-mell set of materials could instantiate the complex dynamics on which cognition depends.

Thus, by making recourse to recent experimental, material, and theoretical developments in active matter physics and soft robotics, in this paper we argue that the separation of non-mental, living "hardware" and cognitive "software" has grown increasingly suspect—and that a harder pivot toward the materiality of the body and cognition is now needed. In other words, what we are claiming is not (or not simply) that the body matters to cognition (a view that tacitly supposes the body in service of a more or less higher-order, more or less unified, cognitive subject), but rather that the body itself—at varying levels of organization—exhibits cognitive capacities through and through: from cellular activities entrained to regulating morphology, development, and intercellular communication; to tissue complexes and system functioning; through to more baroque appearances of cognitive sophistication encapsulated in cephalopod, arthropod, avian, and mammalian brains—Darwin's "endless forms most beautiful" (*Origin of Species*). In a slogan expressed elsewhere (Levin, 2019, 2020; Levin and Dennett, 2020), this is cognition *all the way down*, not just proprietary to a unified subject. Making sense of the theoretical commitment behind this claim and how it contributes to the development of intelligent machines is the main goal of our paper. It is thus worth clarifying at the outset that our discussion of robotics pertains to what we could consider Autonomous Robots (AR), i.e., autonomous embodied systems capable of recursive self-organization, goal-directedness, and agency—the ability to flexibly and actively select goals relative to its "existential needs" (Froese, 2016; Egbert, 2022) and remain the kind of system it is (Man and Damasio, 2019). The key here, as we see it, is to understand how the matter matters to being this kind of system.

The picture we would like to work against is one of neurocentrism that cleaves neuronal (and cognitive) activity from the living, developmental, and morphogenetic processes for which nervous systems originally evolved (see Lyon, 2006; Van Duijn et al., 2006; Keijzer et al., 2013; Newman, 2016, 2019, 2022; Levin, 2019, 2020; Fields and Levin, 2020; Sims, 2020, 2021; Fields et al., 2021; Jekely, 2021; Lyon et al., 2021; Wan and Jekely, 2021). It is this sense in which we think the tacit commitments of MR—the in principle cleaving of active, living processes and cognitive ones—deserve a reconsideration. As Peter Godfrey-Smith remarks, philosophers and cognitive scientists tend to operate with a "picture in which living activity is a kind of non-mental substrate, and then evolution lays a computer—the nervous system—on top of the merely living, after which cognition and subjective experience result" (Godfrey-Smith, 2016a: 496). This can be seen in the very structure of the cognitive sciences and its lack of (explicit) emphasis on the life sciences. That is, while biological perspectives have influenced theorising about the mind [e.g., autopoiesis (Varela et al., 1993; Weber and Varela, 2002) and enactivism (Di Paolo et al., 2017)], they have not furnished real competitive alternatives to more mainstream cognitivism and computationalism [see Meyer and Brancazio (2021) for an insightful discussion]. To this day, it is common to see neurons and the brain—the "stuff" of cognition—almost wholly abstracted from the life processes in which they are embedded.

Here, we hope to cast doubt on the (un)happy divorce between material and cognitive processes by suggesting that looking toward recent developments in soft robotics (Man and Damasio, 2019; Blackiston et al., 2021; Bongard and Levin, 2021; Kaspar et al., 2021; Kriegman et al., 2021), active matter physics (Hanczyc and Ikegami, 2010; Needleman and Dogic, 2017; McGivern, 2020; Egbert, 2021), and basal cognition research (Lyon, 2006, 2015; Van Duijn et al., 2006; Newman, 2016, 2019, 2020, 2021; Levin, 2019, 2020; Bechtel and Bich, 2021; Lyon et al., 2021) complicates any cleaving of cognition from its living, material context. In light of recent empirical advancements, we argue now is a good time to revisit our philosophical assumptions regarding the MR of the cognitive and suggest that a more promising path in the development of AR and Artificial Intelligence (AI) is to take the materiality of cognition *more*, not less, seriously—a position explicitly disallowed in standard philosophical positions on MR. Our argument thus consists of two interlocked moves: first, we identify a set of assumptions that structure the debate on MR and that generate strong intuitions regarding the mental-physical interaction that have historically discouraged taking the materiality of cognition seriously. Second, we propose a path to AR that explores a more thoroughgoing, "radically embodied" approach: one that does not see the body as a "non-mental" substrate on top of which cognitive software (the nervous system) is placed, but instead

depicts cognition as a more fundamental feature of cellular (read: living) activity and self-organizing processes in far-from-equilibrium conditions that are then *scaled up* in appropriate ways to arrive at more sophisticated multicellular animals.

At this point, it is worth being explicit about three things. First, we draw a strong connection between living and cognitive processes—consistent with much of the literature on the so-called life-mind continuity thesis (Maturana and Varela, 1980; Thompson, 2010; Sims, 2021). *Prima facie*, this would seem to undermine our goal of constructing AR, as it would suggest some of the prototypical cognitive behaviours we see in certain soft-bodied robots and active material systems (examined in Section Active matter and soft robotics: Novel approaches to cognition and embodied robotics) cannot qualify as such due to their non-living nature. We believe this problem can be ameliorated, however, by adopting a conception of cognition that depicts the living and developmental side of the process as a more general feature of self-organizing systems in far-from-equilibrium thermodynamic states that must act in a denumerable set of ways to remain the kind of system it is. Simply put: we accept here a view of "life" which does not presuppose particular material foundations (e.g., carbon-based), but rather takes it to be an organizational feature (cf. Moreno and Mossio, 2015). Under this view, then, cognition can be seen as tailored for the homeostatic processes that underpin goal-directed, autonomous, and agentic behaviour (see Pezzulo et al., 2015), and a non-living system would fall closer to the cognitive the more it embodies such dynamics. This brings us to the second consideration, namely, that the lynchpin for our discussion of cognition turns around the notion of "existential needs" and can be explicated, following Lyon et al. (2021), in relation to the set of sensory and information processing mechanisms organisms have for familiarising themselves, valuing, and interacting actively with the environment in order to meet the existential needs of survival, persistence, growth, and reproduction. In the literature, this is often called *basal* cognition, as it refers to a set of mechanisms and capacities with highly ancient, highly distributed origins. We earmark this for now and return to it in Section How fine-grained functional details matter to cognition for a more nuanced discussion. Lastly, it is important to clarify the scope of the present paper. While we engage a diverse range of empirical literature—from active matter physics to soft robotics—we ultimately position the paper at a theoretical level that targets the metatheoretical assumptions that scaffold debates on cognition and mind in Robotics and AI. Stated differently, what we are trying to target is a certain set of assumptions and presuppositions that have historically dominated this field, which complicate taking the materiality of embodiment further than is currently being explored in emerging areas of the life and mind sciences. However, while the present piece is considered theoretical, we believe it encourages actionable and implementable possibilities for creating autonomous systems by incorporating elements of self-organizing dynamical systems (Pfeifer et al., 2007)—as is increasingly explored within the domain of active matter physics and soft robotics.

Our focus on existential needs depends on recent research advocating for taking the materiality of our embodiment further than mainstream embodied cognition has commonly done (cf. Müller and Hoffman, 2017). We thus place a premium on the very processes, goals, and demands of a living body that are normally elided from more theoretical meditations on the cognitive. A similar approach has been proposed by Man and Damasio (2019) who suggest we transition away from the hard parts that typify traditional roboticist approaches to fragile, vulnerable, and soft materials characteristic of organismic embodiment. The fundamental innovation introduces homeostasis and risk-to-self as the warp and weft of cognitive embodiment: "These machines [our AR] have physical constructions—bodies—that must be maintained within a narrow range of viability states… Rather than up-armouring or adding raw processing power to achieve resilience, we begin the design of these robots by, paradoxically, introducing vulnerability" (Man and Damasio, 2019: 449). Indeed, similar to Man and Damasio, we believe a shift from embodied (*simpliciter*) AI to homeostatic and precarity driven AI is the key requirement for the coming generations of AR. This puts more emphasis on the material *processes* and material situation than simply focusing on embodiment full stop. The second notion we depend on has already been mentioned: that of precarity. Tom Froese has argued that the nature of our embodied precariousness (risk-to-self) is essential for agency and the problem of meaning [we might call this a species of the frame problem (McCarthy and Hayes, 1969): why would an artificial agent come to care about its existence and actions on which it depends?]. He writes, "The precariousness that is intrinsic to all organismic, and therefore also of all mental, existence is the original reason why things matter to that individual being" (Froese, 2016: 34). That is, organisms are cognitive agents with meaningful engagements with the world because, and not in spite of, their fundamentally precarious nature. Importantly, this can be also expressed in terms of values and value-realising, which some believe to be the main force driving and organizing action in cognitive agents (cf. Hodges and Baron, 1992; Hodges and Raczaszek-Leonardi, 2021). Precarity is the minimal form of valence, hence enabling cognition and agency (cf. Lyon and Kuchling, 2021). Thus, if "the problem with AI", as John Haugeland famously put it, "is that it doesn't give a damn" (Haugeland, 1998), then we explore how an active matter lens focusing on specific material reconfigurations that enable systems to maintain themselves in far-from-equilibrium conditions can make headway on this most defining of problems for computer science: autonomous robots that might one day give a damn.

The structure of this paper will be as follows. In Section Traditional vs. fine-grained functionalism we briefly overview some of the theoretical and philosophical literature on MR before suggesting that the tenability of its more radical iterations turns on a few key (and, as we would like to suggest, misguided) assumptions that are in need of a rethink in light of recent theoretical and empirical advancements. Instead, we aim to put forward a version of MR that *both* takes the materiality of cognition seriously *and* allows for cognition to be instantiable in alternative media. Section Active matter and soft robotics: Novel approaches to cognition and embodied robotics turns toward the state-of-the-art research to suggest that the domains of active matter physics and soft robotics encourage us to reformulate how we understand the mental-physical interaction. Finally, in Section How fine-grained functional details matter to cognition we defend a "cognition all the way down" approach to the development of bio-inspired AI.

## Traditional vs. fine-grained functionalism

Before putting forward our positive proposal, we need to highlight some of the deeply entrenched philosophical assumptions of the current programme of artificial intelligence that we believe to be detrimental.

The methodology of contemporary AI research is built on top of the philosophical programme of functionalism in philosophy of mind. Functionalism was developed in the second half of the 20th century in response to the issues surrounding the physicalist mind-brain identity theories dominant at the time. In the late 1960s Hilary Putnam advanced a novel line of thought, which sought to establish that mental states (and properties) are *functional* states. From the onset of this view functions, understood as causal mappings between sensory inputs, other internal states, and behavioural outputs (Levin, 2021), were defined in broadly computational terms. This allowed philosophers to disentangle cognition from its neurophysiological, material basis and argue that psychological (which was the main term used for what we call "cognition") processes are "General" (see Polger and Shapiro, 2016: 15), i.e., shared across species, and in fact that psychological functions can be realised by entirely distinct types of systems—not only differently organized animal brains but also a variety of non-biological systems. A special case of interest concerned digital computers, which seem under this view to be well suited for realising psychological processes. This is the idea that has come to be known as the claim of "multiple realisability" (MR) of psychological states.

As Chirimuuta (2018) observes, functionalist theory of mind and the concept of multiple realisability hold a unique status in philosophy as views to which a "near majority of philosophers have subscribed to, and for more than one generation". However, in an important book, Polger and Shapiro (2016) argue that the view of multiple realisability is not, in fact, borne out by the empirical evidence accrued over time. The main thrust of Polger and Shapiro's arguments is aimed at the tenuous distinction between inherent superficial variation in the biological world and deeper differences which are in fact responsible for the multiple realisation of cognitive functions. This in fact turns out to be damning regardless of whether one assumes that mental states are multiply realisable functions (the ontological, objective stance) or whether one argues that they can be explained as multiply realisable functions (the epistemological, subjective view).[1] Their points are targetted at what can be called "MR 1.0" (Chirimuuta, 2018) and, in result, call for a rejection of traditional functionalism. MR 1.0 is, Polger and Shapiro argue, untenable given contemporary empirical evidence.

The functionalist account has suffered from other important theoretical criticisms as well, among which we may highlight the dual objection that functionalism is either (1) too liberal under one reading or else (2) too chauvinistic under another; and, what is more, there are no other interpretations available to it. Following the first option, under which functionalism fails to specify any restriction on the domain of physical systems, it will assign mindedness to entities that should not be viewed as minded. In fact, an important argument in this vein comes from Putnam himself, who later in his life rejected the computational theory of mind. Putnam (1988) proves the theorem that "[e]very ordinary open system is a realisation of every abstract finite automaton", which would lead to an uncontrolled expansion of systems that we should consider as realising cognitive functions—contrary to our experience with the world. The opposite argument has been initially suggested by Ned Block. Block (1978) argues that any version of functionalism that avoids liberalism by opting for some set of physical specifications falls into biological chauvinism and hence denies mentality to creatures that we would ordinarily consider as such. His reason for the claim is the thought that one could always conceive of some system that would fail to meet the physical constraints and yet intuitively seem to possess psychological states.

However, despite the problems with traditional functionalism, the conviction that multiple realisability is an important feature of the cognitive remains widespread among researchers. In fact, it plays a significant role not only in

1 We are grateful to an anonymous Reviewer for pointing out that the ontological and epistemological interpretations of the claims of functionalism and MR are often conflated and prompting us to clarify our position. One could take this argument further, along the lines of what the Reviewer suggests, to point out that science is not in the business of making ontological claims, except only for practical purposes. While some of the authors of this paper are sympathetic to this view, we believe our discussion of traditional vs. fine-grained functionalism is appropriate regardless of one's views on metaphysics of science.

the study of cognition but in the life sciences at large. This led Chirimuuta to propose that instead of rejecting MR altogether, we need to carefully update this notion to account for the role that functional thinking beyond traditional functionalism plays in biology.

Chirimuuta's conception of "MR 2.0" is grounded in observations of how the consideration that certain biological processes are best described as functions that can in principle be multiply realised is an important assumption that allows scientists to make decisions with regard to what physical properties can be safely ignored in their experiments, reducing the complexity of the problem to be studied. This conception of MR, or so Chirimuuta suggests, allows us to maintain that "[i]t can both be true that the material from which the nervous system is built (i.e., living, metabolizing cells) is crucial to their function *and* that those functions are multiply realised" (Chirimuuta, 2018: 411). In particular, this view lets us appreciate that the Heraclitean nature of biological material—its ability to preserve integrity through "continual turnover of matter and energy" (Chirimuuta, 2018: 411)—is crucial for understanding the functioning of cognition [as Godfrey-Smith (2016a) also argues], but at the same time a roadblock to the success of purely reductive methodologies, as this constant shifting obfuscates functionally relevant patterns which occur at a meso-level of description.

This updated view of MR is, in fact, compatible with the idea of "fine-grained functionalism" advanced by Godfrey-Smith (2016a). While Godfrey-Smith explicitly rejects the idea of MR, his arguments are aimed at the concept of MR 1.0. The traditional functionalist account, which builds on the older concept of MR, can be characterised, in Godfrey-Smith's terms, as "coarse-grained functionalism." The two are distinguished by the level of organization that they focus on in identifying and characterising the relevant states and processes. Godfrey-Smith accepts a multi-layered view of reality and concedes that "[t]here are reasonable coarse-grained senses of 'learn' and 'perceive' in which anything with the right coarse-grained functional profile, including a robot, does learn and perceive" (Godfrey-Smith, 2016a: 501). However, he moves on to argue, the systems that we know to be cognitive and proto-cognitive, i.e., a variety of organisms, have an entirely different fine-grained make-up. Not only is it important that in living systems "the information processing side of its activity is integrated with the metabolic side" (Godfrey-Smith, 2016a: 502) but the small spatio-temporal scale at which cellular metabolism occurs has several unique characteristics. In particular, the cells are full of a *molecular storm* with "unending spontaneous motion [...]. Larger molecules rearrange themselves spontaneously and vibrate, and everything is bombarded by water molecules, with any larger molecule being hit by a water molecule trillions of times per second." The ubiquitous electrical charge is just one form of energy present, as chemical, kinetic, and electrostatic energy are constantly transduced into one another. Each part of the cell is subject to forces stronger than it can exert and causality is best perceived as "biasing tendencies in the storm, nudging random walks in useful directions" (Godfrey-Smith, 2016a: 485–487). Cellular metabolism arises from this material volatility and constant flux and, as Godfrey-Smith underscores, principles governing it remain crucial for the processes that constitute cognition, due to their co-evolution.

While the exact dependence of the mind on these low-level processes remains an open question, Godfrey-Smith argues that fine-grained functionalism can account for the failure of traditional functionalist approaches to understanding and engineering minds. Consider a machine—a computer—or a cyborg; even if it has similar coarse-grained functions, it will be lacking the fine-grained functions which depend on the living (i.e., far from thermodynamic equilibrium) organization of biological organisms. It may be capable of "sensing" or "learning", but these terms, or so Godfrey-Smith argues, are broad and coarse-grained, such that they do not rely on a similarity between the fine-grained functional profiles of sensing machines and sensing humans. Reality is multi-scaled and so focusing only on the scale of such coarse-grained properties will not yield the kind of understanding of cognitive processes we need to build intelligent artificial machines.

For Godfrey-Smith this view leads to a rejection of MR altogether but that is the case only for the traditional conception we call "MR 1.0". "The finer-grained features are not merely ways of realising the cognitive profile of the system. They matter in ways that can independently be identified as cognitively important", he argues (Godfrey-Smith, 2016a: 503). He indicates the inherent historicity of neurons—the change in their functional profile resulting from their own activity—as an example. This argument paves the way for Chirimuuta's upgraded notion of MR 2.0, which would hold that fine-grained functions and the material basis of cognition need to be centred in their own right, but could still, at least in principle, be multiply realised. Interestingly, a related point has in fact been a source of criticism for Godfrey-Smith's view raised by Brunet and Halina (2020), who discuss the existence of molecular machines—computers which preserve some of the low-level characteristics indicated by Godfrey-Smith—as an argument for the possibility of developing artificial sentience, which Godfrey-Smith appears to deny. However, given the discernible compatibility of Chirimuuta's MR 2.0 with Godfrey-Smith's fine-grained functionalism, it is more useful to consider Godfrey-Smith's rejection of contemporary approaches to AI to be concerned solely with their focus on coarse-grained functions.

To this list of grievances with regard to the traditional functionalist assumptions underpinning the current AI frameworks we may add one more, namely, that the coarse-grained functions they try to realise *in silico* are inherently highly complex. These are usually specific to a human way of engaging with the world, loaded with folk-psychological ideas, and disjointed from their evolutionary and developmental

trajectory. In result, they are disconnected from the various scaffolds that biological intelligences use for the same purpose. This means that when trying to implement a particular psychological function in a computer AI researchers face a much more difficult problem than the one that evolution faces. If this consideration is correct, a "molecular computer" of the sort examined by Brunet and Halina (2020) would not be sufficient to be deemed a promising candidate for sentience, as developing the requisite coarse-grained functions on this platform would constitute a similarly difficult problem as in the case of silicon-based computing. This is because Brunet and Halina's view relies on the implicit distinction between the computational "hardware" of molecular computers and cognitive "software" (problematised in the introduction and discussed in greater detail in Section How fine-grained functional details matter to cognition). They are interested in the possibility of designing "universal Brownian circuitry capable of extracting useful computation from nano-scale fluctuations" (Brunet and Halina, 2020: 233) and instantiating cognitive processes on top of this circuitry. But that means that their understanding of the functions of cognition remains coarse-grained and hence disjointed from the properties of fine-grained functions. As a result, an AI researcher working on this platform would face a problem as difficult as when working on computing platforms employing standard, von Neumann architecture. The necessary missing step, consistent with fine-grained functionalism, seems to be the use of competent, intelligent parts in the manner suggested by work within basal cognition (e.g., Levin, 2019). We will explore this view in detail in Section How fine-grained functional details matter to cognition.

It is important to note that while both Godfrey-Smith and Chirimuuta leave their claims about the relevance of materiality for cognition at a pretty abstract and general level, we believe that several interconnected research fields—particularly active matter physics and soft robotics that are the focus of the current paper—allow for substantiating these claims further. Notably, doing so lets us draw some initial hypotheses about what "suitable organization" presumed by fine-grained functionalism could consist in and how metabolism may fit into this picture. We turn to the discussion of these disciplines in the next section.

## Active matter and soft robotics: Novel approaches to cognition and embodied robotics

In the previous section we overviewed some of the contemporary literature on MR, specifically regarding the cognitive. We referenced the fact that some of the basic pretenses of MR 1.0 seem to have grown increasingly suspect in light of empirical advancements in the cognitive and life sciences. Indeed, the crux—for our argument—is the condition that the material configurations instantiating cognition be "suitably organized", a requirement that is a lot more stringent than proponents of MR 1.0 would allow. To this end, we began to suggest that recent developments in the areas of soft robotics and active matter physics hint that, while dimensions and aspects of cognitive systems can be manifested in alternative media, they do so *insofar* as they approximate the organizational, living, and developmental dynamics to organismic cognition—a position that we have called fine-grained functionalism.[2] Thus, this section turns toward the empirical basis for something like fine-grained functionalism to adumbrate how the material out of which embodied agents are constituted is integral to sustaining the self-organizing dynamics on which cognition depends. Our main goal, then, is to suggest how a more thoroughgoing, "radically embodied" approach to AR and AI supplies the requisite tools to advance the field toward intelligent, plastic, and adaptive machines (Man and Damasio, 2019; Pishvar and Harne, 2020; Kaspar et al., 2021).

Before continuing, it is worth anticipating briefly why this approach is, or so we want to suggest, a more thoroughly embodied approach than previous iterations of embodied cognitive science. Consider how multicellular agents are themselves constituted out of highly competent, cognitive units (Baluška and Levin, 2016; Levin, 2019, 2020, 2021; Levin and Dennett, 2020; Lyon et al., 2021). In other words, the cognitive cogency of the higher-level (in this case, multicellular) agent depends on the scaling up (see Section How fine-grained functional details matter to cognition) of the cognitive processes—agency, goal-directedness, decision-making, memory, learning—found in the dynamics of constituent (somatic) cells. As we will see (in Section Soft robotics), individual cells are remarkable structures that, due to their regulatory and organizational dynamics, maintain internal milieu viability and their connectivity with other cells in the extracellular tissue complex with precision and flexibility. Reminiscent of 19$^{th}$ century theories of the "cell state" (Reynolds, 2007), our approach thus positions organismic cognition as emblematic of the homeostatic and self-organizing

2 What we mean by "approximate" here is that such systems need not necessarily meet all the criteria that characterise the organizational, living, and developmental dynamics of organismic cognition, and yet they can still exhibit interesting properties, which allows researchers to home in on the causal structure of cognition. In a sense, this is reminiscent of the role of idealisations in scientific modelling (see Potochnik, 2017; except that here such an approximating physical system would be considered both the "model" and the phenomenon model). This is not an exhaustive characterisation, partly because of what an anonymous reviewer has pointed out, namely, that some of the difficulties involved in this project stem from not having any generally accepted cases of non-biological cognition–and even cases of non-human cognition are deeply contested. However, this issue does not detract from the core arguments of the paper and so a more in-depth discussion will be left for future work.

processes that typify all living units. Or, as Man and Damasio put it, "high-level cognition [is] an outgrowth of resources that originated to solve the ancient biological problem of homeostasis" (Man and Damasio, 2019: 447)—hence, cognition all the way down.[3]

As it happens, the building of higher-level cognitive agents out of progressively smaller—but still cognitive—units and parts is precisely the perspective being taken up in the domain of soft robotics and synthetic biology. Ebrahimkhani and Levin (2021) provide a flavour of this style of argumentation:

> "One feature of bioengineering at the meso-scale that is unique… is the fact that bioengineers build out of parts that are themselves highly competent, for example, cells that have their own internal homeostatic and signalling systems. Thus, the experiments that are done with biological parts have the potential to help understand how swarm intelligence plays out at the tissue level to solve morphogenetic problems. Such advances… act as an inspiration for novel architectures in machine learning, artificial intelligence, and resilient autonomous swarm robotics."

Indeed, this bio-inspired approach feeds well into current ambitions of developing AR and unconventional computing platforms (e.g., Jones, 2015). The key is how the above sciences emphasise the importance of an *active matter* approach. Rather than the inert, hard, and passive parts traditionally used in robotics, active matter approaches indicate how the very materiality of the system can perform complex feats that obviate the need for overarching or centralised control (Bechtel and Bich, 2021; Kaspar et al., 2021). In what follows, we survey the fields of active matter physics and soft robotics to then return in Section How fine-grained functional details matter to cognition to our fine-grained functionalist take on how the matter matters for life and cognition. By now, it should be clear that in arguing this position we are not being substantialists: it is not this or that type of matter (say, carbon) that is important, but the matter insofar as it can sustain organizational complexity of the right sort.

## Active matter

Active matter physics (AMP) is a vibrant field of research that has received significant attention in recent decades (see Baez, 2021). Theoretically, it sits at the intersection of physics and biology and deals with materials and material systems that are intrinsically out of thermodynamic equilibrium. Some examples of these are field-responsive matter, hydrogels, and piezoelectric materials, while active matter systems are those that harness properties of such materials to drive their distinctive non-equilibrium behaviour. These include the actomyosin cytoskeleton (Needleman and Dogic, 2017; Jülicher et al., 2018), cellular activities (Fodor and Marchetti, 2018), swarming behaviour (biofilms, multicellular bodies: Wioland et al., 2016; Kempf et al., 2019), and even macroscale organizations such as avian murmurations and herds of animals (Cichos et al., 2020). Although this might appear to be a heterogeneous set, these systems exhibit the broad commonality that their individual units (motor proteins, cells, individual organisms) are themselves highly competent, active contributors to group dynamics (Needleman and Dogic, 2017).

For example, it is increasingly common to view multicellular bodies as a kind of swarm behaviour (Arias Del Angel et al., 2020), which depends on the intrinsically active nature of constituent cells. Indeed, Arias Del Angel et al. (2020) have commented on how facultative multicellularity in both protists and prokaryotes depends on active, field-responsive, and internally driven physical processes of constituent parts, remarking that the overall organismic form hinges on the interplay of the inherent physical properties and agent-like competency of cells making decisions in a context-sensitive and flexible manner. In contrast, then, to passive systems (e.g., the Rayleigh-Bénard cell) that receive energy exogenously at a boundary condition, active matter systems—of which organisms and certain designed systems are paragons—themselves consist of units that are internally driven (Batterman, 2021). Crucially, Needleman and Dogic (2017) remark that active units are capable of *self-organization*, whereas passive units can only *self-assemble*. In the context of being "suitably organized" what we see is that not any back-of-the-envelope set of materials can sustain the organization dynamics on which life and, we add, cognitive processes depend—instead, to be suitably organized one must have self-organization, and it is here that an active matter approach is most pertinent.

Seeing how the dichotomy of active and passive structures underpins much of the literature in AMP, it is worth explicating further what marks out the former exactly. In their influential review of AMP, Marchetti et al. (2013) write that active matter systems consist of the following features:

> "They are composed of self-driven units… each capable of converting stored or ambient free energy into systematic movement. The interaction of the active particles with each other, and with the medium they live in, give rise to highly correlated collective motion and mechanical stress. Active particles are generally elongated and their direction of self-propulsion is set by their own anisotropy rather than fixed by an external field (Marchetti et al., 2013: 1144)."

3 As an anonymous Reviewer pointed out, this approach is also supported by the claims made by the proponents of the Free Energy Principle framework (e.g., Friston, 2019). Exploring this topic in sufficient depth, however, would require a separate paper, and hence we have to refrain from expanding on this connection here.

The key distinction we wish to draw out here is that being an active matter system relies on two features: (i) the energetic nature of the constituent units (actively converting ambient energy as opposed to being driven solely by energetic contributions at an external boundary conditions) and (ii) their inherent shape (anisotropy) influencing the systematicity or directionality of how energy is used—"geometry of [its] interface shape can control sensitivity to the environment" (Hanczyc and Ikegami, 2010). The example of the Rayleigh-Bénard cell will help draw out the important difference between the two types of material configurations (i.e., passive vs. active).

Rayleigh-Benard cells are a paradigmatic case of non-equilibrium activity. These are familiar to anyone who has ever added cool oil or water to an evenly heated pan. The sudden encounter of the droplets with a hot surface drives a phase transition that constrains the activity of individual water molecules. The result is a highly ordered hexagonal structure that is continuously sustained so long as energy input is consistent. Some commentators remark on how structures like the Rayleigh-Bénard cell represent the precursor dynamics from which goal-directedness endogenously emerges (Juarrero, 2015), and they are hence common reference points in discussions of emergence, agency, and goal-directedness (cf. Moreno and Mossio, 2015).

Despite its relevance as a model for far-from-equilibrium processes, the Rayleigh-Bénard cell does not qualify as an example of an active matter system. The reason for this has already been suggested above, but Needleman and Dogic make it clear: "Rayleigh-Bénard patterns are non-equilibrium dissipative structures, but each convection roll is composed of passive molecules, and the entire system is driven away from equilibrium by energy provided through an external macroscopic boundary" (Needleman and Dogic, 2017: 1-2). They meet neither requirement (i), as they are composed of energetically passive molecules, nor requirement (ii), as their shape is a result of the motion guided by an external energy gradient, in no way dependent on inherent properties of the medium itself—which, without influence, would immediately relax into an amorphous shape as normally water molecules on a pan tend to do. Contrastively, active matter physics addresses the activity of thousands of nanoscale molecular motors that interact to create mesoscale, self-organizing structures. Common examples span the living and non-living domains, including model systems such as self-propelled oil droplets (Hanczyc and Ikegami, 2010; Hanczyc, 2011; Cejkova et al., 2014), active microtubule networks (Sanchez et al., 2012), cytoplasmic flow (Mogilner and Manhart, 2018), and the eukaryotic cytoskeleton (Brugues and Needleman, 2014). More recently, active materials have been exploited in soft robotics (Ebrahimkhani and Levin, 2021), computer science (Jones, 2015), and AI (Kaspar et al., 2021) as a way to overcome the many resource constraints that have long plagued the fields. The key point can be expressed as follows: "The cellular cytoskeleton, cells, and entire tissues [as exemplary active materials] are driven away from equilibrium by the continuous motion of thousands of constituent nanoscale molecular motors, protein-based machines that transform chemical energy into mechanical motion" (Needleman and Dogic, 2017: 2). Intriguingly, this is a point that has echoes in Section Traditional vs. fine-grained functionalism in our discussion of fine-grained functionalism and the relevance of spatial *scale*: "Metabolic processes in cells occur at a specific spatial scale, the scale measured in nanometres… In that context and at that scale, matter behaves differently than how it behaves elsewhere. … There is unending spontaneous motion that does not need to be powered by anything external" (Godfrey-Smith, 2016a: 485). At larger, more coarse-grained scales, these complex and systematic processes do not occur. Already, then, we come to see how fine-grained structural details matter for sustaining self-organizing dynamics at a wider variety of scales.

Of course, what is central to the discussion of fine-grained functionalism is the connection between these active material processes and prototypical instances of cognition, such as goal-directedness, memory, learning, agency, systematic directionality, and so on. As it happens, recent work on active materials has begun to show the variety of ways in which some individual—and sometimes multiple—capacities are present in non-living systems, a discovery that has led some to speculate that AMP is revealing not only the physics of life (Popkin, 2016), but the physics of cognition as well (McGivern, 2020). To wrap up the discussion of AMP, then, we make a more direct connection to work on basal cognition and the concept of existential needs introduced above.

Capacities of non-living active matter systems that have been particularly illuminating are those of autonomous movement, environmental sensing, coordinated action, and problem solving (McGivern, 2020). The ability to accomplish these feats importantly depends on the material situation of both the system in question (swarming nanobots, self-propelled oil droplets) and the environment where it finds itself. In self-propelled oil droplets, for example, researchers introduce internal convection currents that create a bifurcation between systematic internal activity and its viscous medium. The droplet's movement is driven by a convective flow that has an uneven influence on the inside of the droplet, which helps create a feedback system between its internal dynamics and the medium external to it (so-called Marangoni flows; see Hanczyc and Ikegami, 2010 for a review). Although this is a simple system, Hanczyc and Ikegami (2010) suggest that it serves as a model system for understanding the origins of chemotaxis in unicellular organisms, as the droplet must continuously navigate gradients to find the chemicals that sustain its internally driven dynamics: "The system becomes sustainable by circulating the reactants and products effectively as organized by the convective flow" (Hanczyc and Ikegami, 2010: 236). As we can see, it meets both conditions of active matter listed above: (i) its constituent particles are internally

energetically driven in that they tap into reservoirs of (Gibbs) free energy available within the system, and (ii) its droplet shape results from the inherent properties of the oil (its viscosity and surface tension) and its relation to the medium in which it is embedded; moreover, the geometric configuration of the oil droplet actively contributes to the distinctive capacities it exhibits. This is an intriguing model system for our purposes, as it places a premium on active materials and constituent units that spontaneously self-organize and, given the right guidance and influence from designed experimental parameters, can sustain itself for significant periods of time. Although not elaborated here, the case of oil-droplets also underlines the way in which the inherent shape ("geometry-induced fluctuations"; Hanczyc and Ikegami, 2010) of an active unit determines locomotion and, in bacteria, chemotaxis. Moreover, a mechanical pushing of the cytoplasmic sol of a cell (as in the social amoeba *Dictyostelium*) elicits directional and coordinated motion (see Dalous et al., 2008; Boussard et al., 2021). Thus, the properties exhibited in active matter systems, such as oil droplets, highlight the material basis for capacities found throughout the living domain.

AMP shifts our focus on the study and development of minimally cognitive systems (that is, systems that exhibit prototypical features of cognition such as directional locomotion, memory, or learning) in two important ways. First, it does not aim to replicate paradigmatically intelligent behaviour modelled on human activity (playing chess, say) and instead emphasises environmentally embedded behaviour with wide distribution in the natural world (McGivern, 2020). Secondly, and perhaps most centrally for our argument on MR 2.0, "work on active materials is not specifically aimed at computational characterisations of behaviour" (McGivern, 2020: 442), i.e., it does not rely on coarse-grained functions of the medium of interest, but rather builds on simpler, well-established principles from areas such as condensed matter physics, building a bottom-up description of activities of systems of interest. Work within AMP, then, demonstrates how harnessing the physical processes and active materials that underlie organismic behaviour contributes to and mediates the cognitive sophistication we find in the biological domain—suggestive of how such principles can, and are, being exploited in the domain of artificial and designed soft robotic systems, which we turn to shortly.

Before continuing, however, it is worth dwelling on the aspect of AMP that we see as central to the discussion of cognition that forms the remainder of this paper. Recall from the introduction that our understanding of cognition revolves around the fulcrum of existential needs and how capacities such as agency, goal-directedness, and self-maintenance are the basis for further cognitive sophistication. Matthew Egbert has recently argued that non-biological model systems—such as our humble oil droplet—serve as ideal testbeds for exploring the material and thermodynamic basis of these existential needs: "conditions that must be met for [that system] to persist and… behave in ways that satisfy those needs" (Egbert, 2021: 5).

There are two ways to understand existential needs vis-à-vis any object or system, the first rather banal and the second more critical for the kinds of systems we explore in this paper. The first is the sense in which, trivially, any object must have existential needs to be what it is. A table cannot be heated above a certain temperature or subject to a certain amount of pressure and still remain a table. But, and this is the more important point, there are crucial differences between what is required of garden-variety non-dissipative objects like rocks, tables, and chairs to be what they are and self-organizing, self-maintaining dissipative systems in far-from-equilibrium conditions. The difference in existential needs for the two types of systems is captured as follows:

> "[Non-dissipative entities] are merely passively stable, whereas dissipative structures are constantly falling apart and yet persist thanks to processes of repair, replacement, or reconstruction. This means that existence for passively stable entities is the absence of a destructive event. In contrast, for dissipative structures, existing is a process—and a process that must continue for the system to persist (Egbert, 2021: 5)."

Importantly, processes have quantifiable and measurable rates that open dissipative structures to a study of how viable such a system is, that is, how well it persists despite the tendency to degrade. As Egbert notes, there is no equivalent measurement for passively stable systems: their existence is not a process and does not require the same set of behaviours and activities that active matter systems engage in. We can therefore agree with Godfrey-Smith when he writes "macroscopic machines provide a poor model for the material basis of living activity and for the material basis of mental activity in living beings like us" (Godfrey-Smith, 2016a: 489).

Our discussion of AMP furnishes one strand of the argument for fine-grained functionalism, namely, that the fine-grained material and thermodynamic details of living systems matter a great deal more than common assumptions on the MR of the cognitive might *prima facie* suggest. Indeed, organisms are subject to what physicists call the "tyranny of scales"—they are sensitive to, and influenced at, every order of scale, from the nanoscale to the mesoscale, and for multicellular agents like ourselves, the macroscopic scale. These are highly sensitive coordinated structures, and there is no non-arbitrary point below which the physics no longer matters to manifesting the distinctive cognitive capacities that contribute to a living system's survival. Although our discussion of cognition has been minimal in this section, we turn now to soft robotics to see how these insights are being actively taken up in designing intelligent synthetic machines.

## Soft robotics

Soft robotics is a sub-discipline of robotics and artificial intelligence that explores how intelligent, adaptive, and plastic behaviour emerges out of the inherently active, precarious, and soft parts that constitute such systems. It is a discipline that examines how to construct systems that exploit the physical laws and tendencies at play at every level of scale. In other words, it investigates how organisms are embedded and subjected to a "tyranny of scales" that must be accommodated and exploited to meet their existential needs (Ebrahimkhani and Levin, 2021) and how one may apply these insights to the creation of intelligent machines.

Tellingly, Shah et al. remark that the inspiration for soft bodied robots comes from the highly integrated nature of biological cognition: "In these integrated living systems, intelligence, memory, learning, behaviour, and body structure are all intertwined and emerge from the multiscale dynamics of the same robust and highly fault-tolerant medium" (Shah et al., 2021: 1). This is put in contrast with the standard hard (and passive) components that constitute more standard roboticist approaches. Standard approaches have had some success in the form of modular parts that can be added or taken away depending on the task (such as passive conforming grippers and certain algorithms that can re-adapt to distinct tasks). But even in cases where these techniques might achieve certain adaptive ends, they "operate under the assumption that the robot's body is only reconfigured or reshaped due to external forces, and do not explore the possibility of synthetic machines that actively grow, regenerate, deform, or otherwise change the resting shape of their constituent components" (Shah et al., 2021: 2). Contrastively, as we saw above, the field of AMP begins to highlight the way in which organism morphologies and bodies are inherently active structures that respond proactively to changing environmental situations—a form of adaptiveness that depends crucially on the highly active processes that comprise cellular structures, multicellular integration, and cognitive capacities such as goal-directedness and agency. This picks up on a point made earlier: a key feature of organismic cognition and an insight that has been actively taken up in soft robotics is that higher-level cognition relies on constituent parts that are themselves *highly* competent.

Here, we focus on how the concept of existential needs, raised in the introduction, is critical for the creation of artificial machines capable of autonomously selecting actions required for self-maintenance. In other words, in contrast to passively maintained robots that must be externally guided and directed toward goals, tasks, or functions, it is suggested that (i) the inherent vulnerability of soft embodiment coupled with (ii) thermodynamic processes that are required to maintain a system in such a state would endow an artificial agent with the kind of autonomous self-maintenance and self-organization that are important for cognition (cf. Bickhard, 1993). Only then would these designed systems have real "skin in the game" (cf. Bongard and Levin, 2021). To put the matter differently, to design machines capable of autonomous decision-making, behaviours must have *consequences* for how the system can and should act in the world. We have introduced this idea previously in terms of precarity and risk-to-self (in Section Introduction), and with the analysis of active matter above we may further specify the details of what precarity would mean for a machine. Man and Damasio (2019) indicate, in terms intriguingly close to Egbert's paper cited above, that we can understand how feelings emerge from a physiological investigation of life regulation. Feelings, they argue, are not sufficiently approximated by arbitrary reward or loss functions of standard approaches to AI, since the worldly risks and consequences should directly impact the continued existence of the machine. The quality of feeling "is the harbinger of the good or bad outcome relative to survival" (Man and Damasio, 2019: 446). They argue—and we concur—that it is only at the point when the machine can consistently strive for continued existence that true agency may arise.

It is important to note that what we mean here by "life regulation" is not biology-restricted, but rather is the upshot of a far-from-equilibrium system working against the tendency toward dispersal. The suggestion we would like to make here is that to be this kind of system—a system for which there can be situations that matter to it—it must be a self-organizing one constituted by active physical processes inherent to the materiality of the system in question. Soft robotics, then, is in the business of identifying how the material aspects of the body exploit physical laws to expand robot functionality (Shah et al., 2021: 2).

For example, Pishvar and Harne (2020) note that soft robotics incorporates field-responsive smart matter that can induce an internal flux in response to an applied field that tailors material characteristics of the media, influencing its function and behaviour. As they write, "When responding to applied fields, a multitude of internal changes are possible in soft, smart matter" (Pishvar and Harne, 2020: 1). The range of adaptability is thus expanded when one incorporates material properties that are themselves active contributors to overall robot functionality, in contrast to standard hard parts used in robotics, whose adaptability—in the rare cases when they are adaptable—is due to pressure driven forces at an external boundary condition. More recently, Kaspar et al. have argued that "synthetic matter that itself shows basic features of intelligence would constitute an entirely new concept for AI" (Kaspar et al., 2021: 345). They dub this pivot in AI and robotics the "rise of intelligent matter" and reiterate the point that incorporating active materials into AI and robotics programmes would expand robot functionality "far beyond the properties of static matter" (Kaspar et al., 2021: 345). Examples of such smart, active matter systems include artificial thermoregulating skin (Kanao et al., 2015), emergent swarming activity of concerted nanobots (Wu et al., 2021), and xenobots that sit at the intersection of bio- and artificial

engineering (Ebrahimkhani and Levin, 2021; Kriegman et al., 2020, 2021; Blackiston et al., 2021). All authors appear to be in agreement that incorporating the smart and active propensities of soft matter is crucial for achieving autonomous behaviour in the domain of robotics (Pishvar and Harne, 2020; Kaspar et al., 2021)—a suggestion we will turn toward in the next section on the fine-grained functionalism approach to AR.

## Summary

To wrap up briefly, the fields of active matter physics and soft robotics are working lock and step to uncover the diverse functionality, adaptability, and plasticity inherent to certain materials that remain in far-from-equilibrium conditions. They are thus fields that explicitly consider the thermodynamic situation of the system in question. An upshot of this is that not any sort of material can accomplish the diverse behaviour or cognitive sophistication exhibited in the biological domain. In other words, to be "suitably organized" requires attention to the media out of which the system of interest is constituted: the matter matters for cognition and is not a dimension of robot functionality that can be abstracted out. Indeed, the conclusion we wish to draw from this literature is that *more* attention should be paid to the material basis of cognition than is commonly done.

Given the importance of the above two testbeds for exploring the nature of cognition, it is crucial to explicitly articulate the connections that can be drawn between active matter physics and soft robotics. Our reasoning for progressing from the former to the latter is that active matter physics deals with far-from-equilibrium dynamical systems, writ large, and the materials and material constellations that can sustain self-organizing processes on time scales relevant to the biological world. It is precisely these processes that are then exploited and harnessed in guided assembly to arrive at the sophistication we find in the field of soft robotics (Ebrahimkhani and Levin, 2021). While prima facie it might appear that the two fields can work in isolation, Pfeifer et al. suggest why this is not advisable: "it [is] clear that autonomous agents display self-organization and emergence at multiple levels: at the level of induction of sensory stimulation, movement generation, exploitation of morphological and material properties, and interaction between individual modules and entire agents" (Pfeifer et al., 2007: 1088)[4]. In other words, active matter physics in tandem with soft robotics furnishes not only the empirical testbeds for crafting more sophisticated autonomous agents, but also renders tractable notions of emergence and self-organization that are of relevance not only to bodily maintenance and self-preservation, but also cognition and the behaviour required to keep such systems viable. In the following section, then, we dovetail the pieces of the argument laid out thus far to advocate for an emerging approach to the development of AI and AR that stems from fine-grained functionalism suggested in Section Traditional vs. fine-grained functionalism and the results of AMP and soft robotics discussed in the current section; a paradigm that appreciates the importance of the materiality of cognition.

4 An interesting context here, also explored in Pfeifer's work (e.g., Pfeifer and Scheier, 1999), is the increasing popularity of "morphological computation" (see also Müller and Hoffman, 2017). For a discussion of morphological computation in the context of basal cognition (see Rorot, 2022).

# How fine-grained functional details matter to cognition

In this paper, we explore the possibility of developing autonomous robots capable of prototypical forms of valuing and engaging with the world in a goal-directed manner. To this end, we adopted the notion of precarity (i.e., "risk-to-self") and focused on the existential needs of a system to remain in a far-from-equilibrium state. We saw that a step in this direction requires rethinking some of our basic assumptions regarding matter and its relation to cognition, which remain deeply embedded in existing approaches to the mind and brain sciences, as well as in approaches to AI and robotics. Indeed, rather than a "layered-cake" model of levels that renders higher-level cognitive phenomena as resting autonomously from and "on top" of its substrate (i.e., "hardware"), we set out to complicate this picture by emphasising the (bio)physical nature of the structures that support, enable, and implement cognition. What we want to suggest is that more attention must be paid to the *fine-grained* details of the system when understanding and studying cognition—and then recreating it in alternative media. This is the crux of the fine-grained functionalism introduced in Section Traditional vs. fine-grained functionalism. Although it is common to see biologists and philosophers emphasise the "Heraclitean" nature of biological matter and metabolism, we believe recourse to the fields of active matter physics and soft robotics situates fine-grained functionalist views on a sturdied empirical testbed. Thus, the developments in these disciplines enable an exploration and substantiation of claims about precarity and existential needs vis-à-vis cognition, which so far we have explored mostly in the abstract.

In this section, we weave the threads of the argument together to argue that creating AR capable of valence and goal-directedness requires us to think about the organizational and material dynamics of the embodied system in a more thoroughgoing way. In other words, what we are suggesting is not (or not simply) that the body *simpliciter* matters to cognition, as advocates of sensorimotor coordination have long held (see Van Duijn et al., 2006). Rather, we argue for a multiscale account in which cognitive and agent-like

competency is present at nearly every level of a biological heterarchy capable of sustaining the appropriate organization—cells, tissues, networks, the whole organism, and even swarming behaviour of eusocial species (Levin, 2019). In contrast to views that situate cognition as exclusively proprietary to a higher-level organism, we present an account in which the scale and "selfhood" of the cognitive agent are highly malleable, plastic, and vacillatory.

We therefore argue for taking embodied approaches further than is commonly done in two important respects. First, in the multiscale approach just outlined: higher-level systems (organisms or future soft robots) themselves consist of highly active and competent cognitive units. The preservation of cognitive functionality at varying spatiotemporal scales is indeed a crucial aspect of evolvability and robustness in organisms (Levin, 2020). Cognition is then regarded as "an outgrowth of resources that originated to solve the ancient biological problem of homeostasis" (Man and Damasio, 2019: 447). It is construed as an activity of self-organizing and self-maintaining processes fundamental to all living organisms, and that is then appropriately scaled up throughout a biological heterarchy, an idea we explore further below. Second, and here we loop back to fine-grained functionalism, these cognitive capacities depend crucially on the material and (bio)physical details that are standardly abstracted out or relegated to a "hardware" problem. These two central themes are discussed in the remainder of this section.

## Scaling cognition all the way down—and back up again

Organismic embodiment is characterised by highly plastic and adaptive parts responding in a coordinated manner to wider organism-level goals. Empirically, this increasingly seems to rely on essentially cognitive units and intelligent parts—i.e., cells, tissues, networks—acting in a concerted manner that involves an "inter-penetrating, concurrent operation of numerous layers of cognition within the same living system" (Levin, 2021: 4); that is, it involves cognitive units maintaining some degree of flexibility, agency, and goal-directedness that is executed in local and global contexts. In biology, this is often on display in the morphogenetic and ontogenetic unfolding of the organism—a complex process that requires cognition to be tailored to both scale-specific as well as scale-free needs in regulating organism development. Indeed, the ability of organisms to plastically change shape throughout their life cycle is the current envy of soft roboticists, where developing shape-changing robots is a frontier in the field (Shah et al., 2021). Here, we explore the phenomenon of shape-shifting, as it helps illustrate how morphogenetic and homeostatic goals pursued at each level can give rise to robust and flexible behaviour at a variety of scales.

As suggested, an organism's ability to arrive at complex morphogenetic outcomes depends on the interpenetration of these functionalities at a range of spatial and temporal scales, as well as the elasticity and robustness of a (predominantly soft) medium. This contrasts with standard roboticist approaches that incorporate hard parts ["up-armouring", as Man and Damasio (2019) phrase it] and assumes that bodies are only reconfigured due to external forces, effectively neglecting the active and proactive responsiveness that typifies biological media and serves as the foundation for homeostatic, self-organizing processes. Indeed, hard-clad robots might experience change at the movement of a joint, but none within the stuff that constitutes it. Contrastively, biological and soft robotics systems "[change] shape at all relevant scales, globally and locally" (Shah et al., 2021: 10). What is important here is that this process is effectuated through the nested hierarchical structure in which every level can pursue its own local (morphogenetic) goals. The morphogenetic (shape-shifting) outcome of this process is thus not only materially and physically active, but an expression of the cognitive coordination to be found throughout the organism. This is in sharp contrast with current robotics, which largely uses unintelligent parts (Shah et al., 2021).

We can call these systems exemplars of "coordinated structures", following Kelso (2016), which are endogenously self-organized systems determined by their own dynamics. Indeed, a characteristic feature of such structures is that they do not depend on an exogenous "ordering influence" (Kelso, 2016: 491), and some have remarked on how this form of self-organization is the basis for higher-level features of autonomy, agency, and goal-directedness (Juarrero, 2015). Perhaps unsurprisingly, the requirements for coordinated structures are parallel to the defining features of active matter systems, suggestive of the relevant building blocks for engineering artificial analogs that could come to endogenously self-organize to create novel, agentic, and goal-directed structures. It should be clear that the vision of cognition we have in mind here is one in which the system itself has real "skin in the game", and therefore requires this minimum degree of autonomy (Bechtel and Bich, 2021).

Importantly, this focus on internal coordination echoes the prominent view that cognition—as it evolved—initially emerged in the course of evolution for coordinating cellular metabolisms and ultimately multicellular (more minimally, intercellular) activity, particularly spatial and temporal coordination across parts of the system—a position Keijzer et al. call the "Skin Brain thesis" (Keijzer et al., 2013; Jékely et al., 2015). The primacy of internal coordination hints at the profound relevance of electrical oscillatory activity found in biological bodies (cf. SELFOs, see Hanson, 2021), which has been put forward as one of the central mechanisms of synchronization—an important topic that future work will explore in detail. Furthermore, the path of engineering intelligent systems from self-organizing and coordinating intelligent parts, while perhaps not the only one, becomes a clearly feasible approach for researchers, since we see

that in the history of life this trajectory has in fact led to the emergence of cognition.

The shifting local and global coordination of the system (i.e., organism) exemplifies the distributed approach to cognition we have defended herein. It requires that constituent cells and parts maintain certain aspects of cognitive function—memory, learning, agency, decision-making—at least in the service of their own form and function. Levin captures this point well when he writes that "somatic cells did not *lose* their behavioural plasticity … to become parts of metazoan swarms (bodies): they scaled them to enable pursuit of larger goals consisting of creation and upkeep of massively complex anatomies" (Levin, 2019: 5).

Thus, the concept of scaling up, which we have relied on throughout this paper, rests on the idea that multicellularity is itself a highly complex and competitive "environment" that requires local and global morphogenetic goals consisting of trade-offs and top-down constraints between small-scale outcomes and organismic level development. We have already suggested that we can understand this principle of scaling up in terms of internal coordination determined by endogenous dynamics but building on the concept of active matter can help elaborate the idea further.

We can identify an appropriately scaled-up active matter system when two conditions are met: (1) the system is not wholly determined by local causes (see Kelso, 2016), that is, the system behaves in relation to non-local causes; and (2) it exhibits goal-directedness as a coherent unit. Internal coordination results from the conjunction of these conditions and, hence, is inseparable from cognition as we explore it here. Our reliance on active matter is motivated by our fine-grained functionalist claims, which we turn to below.

For now, it is important to emphasise that developmental bioelectricity has been identified as the predominant mechanism behind locally and globally coordinated morphogenetic, developmental, and cognitive outcomes—realising the scaling up of parts into wholes. Bioelectrically coordinated and integrated cells (in the form of an organism or a colony of organisms, as in bacterial biofilms) meet the conditions for a scaled up system laid out above in that (1) the activity of cells often results from information about occurrences happening in a distant part of the integrated system and (2) each part coordinates its actions with others, so that the system as a whole exhibits a consistent behavioural pattern (see Arias Del Angel et al., 2020 for an insightful discussion of this in relation to the social amoeba *Dictyostelium*). Thus, the bioelectrical activity that has often been associated with nervous activity is increasingly seen as an exploitation of highly preserved, ancient, and widely distributed cellular functions and capacities (Prindle et al., 2015)—and we extend the discussion to suggest that this itself hinges on more general properties of cellular, biological, and living material dynamics. This is what in developmental biology (Newman, 2019, 2022, 2021) has been called "biogeneric" processes, indicative of how biological functionality in the service of homeostatic, morphogenetic, and developmental goals is an exploitation of general physical principles of viscoelastic media and oscillatory activity. This again draws a strong connection between the "physics of life" and "physics of cognition" we hinted at above.

Indeed, if—as this research suggests—neurons are specialised exploitations of bioelectrical mechanisms, it is more fitting to see the nervous system as initially (both *evolutionarily* and *ontogenetically*) more a matter of "pulling the organism together" than as specialised for higher-level cognitive functions (cf. Fields et al., 2020). Again, pulling from Levin, we see that "neural networks control the movement of a body in three-dimensional space; this scheme may be an evolutionary exaptation and speed-optimization of a more ancient, slower role of bioelectrical signalling: the movement of body configuration through anatomical morphospace during embryogenesis, repair, and remodelling" (Levin, 2019: 5).

The truly innovative move in the literature on basal cognition (that is, cognition as situated in more "primitive" organisms and cellular activities), then, is the explicit recognition of the cognitive (or proto-cognitive; Godfrey-Smith, 2016a,b, 2017) nature of the activities identified above. Indeed, examples of memory in social bacteria (Dinet et al., 2021), learning in unicellulars and protists (Gershman et al., 2021), decision-making in acellular and cellular slime moulds (Arias Del Angel et al., 2020; Smith-Ferguson and Beekman, 2020; Boussard et al., 2021) have all been identified in non-neural organisms, and it is known that constituent cells in metazoan swarms (i.e., multicellular animals) actively and adaptively manage their morphology, behaviour, and physiology as needed for survival. Again, this is cognition within and throughout biological bodies and therefore is suggestive of a more thoroughly embodied cognition insofar as higher-order organized wholes are dependent on constituent units and parts maintaining, in certain crucial respects, cognitive capacities of far more ancient origins. The ability of evolution and hence organisms to exploit the material properties of cellular processes to yield coordinated wholes is the current envy of soft robotics approaches that still rely on guided self-assembly to arrive at robot functionality (Ebrahimkhani and Levin, 2021).

As we have already hinted in Section Traditional vs. fine-grained functionalism, this differs dramatically from extant AI and robotics approaches that do not avail themselves of such techniques, in that goals can be pursued both at the wider level of the whole organism, at the tissue complex level, and the level of cellular homeostasis and intercellular coordination. Indeed, "the ability of each nested level to have its own local morphogenetic goals… contrasts with today's robots, which are largely made of unintelligent parts" (Shah et al., 2021: 10). To conclude, then, we loop back to our fine-grained functionalism claims to highlight the close imbrication between cognitive capacities of interest in the design of AR (agency, goal-directedness, memory, learning,

self-maintenance) and fine-grained aspects of the materials that should, we suggest, be the focus of current and coming robotics approaches.

## Fine-grained functions of soft materials

Fine-grained functionalism rests on the crucial observation that cognition is not temperature. Allow us to explain. When we approach cognitive systems and try to individuate the functions they perform, we always do so at a particular level of granularity that reflects certain aspects of the observer (their interests, needs, pragmatics, assumptions) and not the cognitive system observed. Philosopher of science Angela Potochnik phrases this as a matter of reading our assumptions of the multilevel nature of the world *into* the phenomena of investigation, imposing an artificial hierarchy on a complex system where there may not be one (cf. Bechtel and Bich, 2021; Potochnik, 2021). Crucially, these different granularities do not simply map onto the distinction between macro- and microstates that some branches of physics find useful. Cognitive functions, such as learning, memory, decision-making, are not macrostates realised by (possibly very different) physical microstates in the way that the same temperature can be realised by various distributions of thermal energy across molecules—even if we may observe these functions in equal part in a variety of natural and engineered systems.

What we mean to say, then, is that non-biological passive materials (in our case, materials that cannot sustain self-organizing dynamics in far-from-equilibrium conditions) will *not do* the same things as soft biological counterparts: "They will be functionally different, not merely different in "hardware" or "make up"' (Godfrey-Smith, 2016a: 501). For functional equivalence, their material structure and organization must occupy specific spatial and temporal scales to endogenously accomplish self-maintenance and self-organization. The main upshot of this view is that fine-grained functional properties of living systems, such as metabolism and recursive self-maintenance, matter quite a deal more than is commonly supposed in debates on the MR of the cognitive. In other words, if cognition depends on suitably organized, endogenously driven internal dynamics of cellular activity and the appropriate scaling up into even smarter wholes, we begin to see the way such details become the foundation of the cognitive—and, by extension, central to the approach to engineering AR that we advocate. There are two considerations with which we would like to conclude.

First, fine-grained functionalism imposes clear constraints on what sorts of materials are capable of instantiating cognition—without falling into the biological chauvinism typically (and erroneously) associated with this type of view. The required platform must be able to sustain self-organizing dynamics in far-from-equilibrium conditions on temporal and spatial scales that make it susceptible to physical forces, constraints, and tendencies that are not found at larger spatial scales—the scale of standard machines to which biological cognition is traditionally compared (Nicholson, 2014). These conditions are met by soft, active materials: a domain of materials science that continues to grow in popularity since its inception in the 1990s. From a physicist's perspective, exemplary soft materials such as "[c]olloids, polymers and surfactants, sometimes also known as 'complex fluids', have one characteristic in common: they involve a mesoscopic length scale between the atomic (∼1 nm) and the bulk (∼1 mm). On this intermediate length scale, one finds structures such as suspended particles/droplets, macromolecular coils, and self-assembled structures such as micelles and bilayers" (Poon, 2000). The ability to self-assemble into vesicles is especially interesting, as, according to some researchers (see Kauffman, 1993), such structures form a necessary step in the emergence of life, since they allow for the prebiotic system enclosed within to control its interactions with molecules in the environment and, in result, to remain at the boundary between subcritical and supracritical behaviour. As stated throughout, we do not preclude the possibility of non-living cognitive systems. Indeed, crucial points of our argument turn on the blurring of cherished distinctions between paradigmatically living and non-living systems.

What we do want to highlight, however, is that a non-living system is cognitive the more it approximates dynamical features of living activity—that such activities (which we broadly associate with self-maintenance and, eventually, homeostasis *via* metabolic activity) are the fount from which higher-level cognition emerges. The overlap in the dynamics between living and non-living systems constrains the types of materials that can enter a concerted organization able to sustain itself recursively and endogenously. While common reference points in philosophical debates on MR 1.0 consist of cognition being instantiated by (*inter alia*) tumbling beer bottles, frenzied radio signalling between denizens of the Chinese nation, and, of course, Swiss cheese, it is clear from what we have argued that these are *not* the kinds of things that can sustain self-organization endogenously. Matter behaves differently at the scale of objects normally invoked to support intuitions on MR, reaffirming the point expressed above that these materials will not do the same thing as the molecular motors, nanoscale molecules, and field-responsive materials we find at length scales well below that of everyday familiarity. Dislodging our intuitions about the MR of the cognitive, and upgrading from MR 1.0 to MR 2.0, enables us to attentively observe the behaviour of matter at nano and mesoscales to more properly assess how proto-cognitive capacities relate to the frenzied activity of fine-grained features of the system—not treating them as "noise" or obfuscating complexity to be abstracted out.

What we do want to highlight, then, is that an active matter approach oriented around soft materials could begin to

approximate these features in non-living systems and media—as the exciting field of soft robotics is beginning to show (Section Soft robotics above). Hence, an attempt to design and engineer an artificial cognitive system in such materials would tend to fall closer to cognition than extant hard-part robotic systems. Soft materials and our active matter lens provide (some of) the resources to better assess the "suitably organized" claim so often made in debates on multiple realisability. As already mentioned, this allows us to resist biological chauvinism worries, while also delimiting the kinds and configurations of systems that can be autonomous cognitive agents, hence neutralising the liberalism charge as well. That is, as stated earlier, the active matter lens allows us to argue *both* that the materiality of cognition matters *and* that the cognitive can be realised in alternative media (Chirimuuta, 2018; Brunet and Halina, 2020).

These considerations bring us to the second important insight granted by the perspective of fine-grained functionalism. It constrains how we should approach the task of engineering AR, defining a feasible—at least so we hope—research strategy. Developing artificial cognition once we have rejected the "hardware/software" distinction renders the concept of Artificial General Intelligence (AGI)—a Holy Grail of present-day AI researchers (explicitly embraced by companies such as OpenAI and DeepMind)—misguided. AGI can be understood as "loosely speaking, AI systems that possess a reasonable degree of self-understanding and autonomous self-control, and have the ability to solve a variety of complex problems in a variety of contexts, and to learn to solve new problems that they didn't know about at the time of their creation" (Goertzel and Pennachin, 2007). The view that emerges from the "cognition all the way down" approach is that cognition is not "General" in this sense—cognition is not a single programme that can be applied to a variety of contexts, in the way that the programme MuZero (Schrittwieser et al., 2020) is able to master a variety of video and traditional games without explicit presentation of the rules. Rather, cognition results from the orchestration of a vast amount of single-purpose, specialised processes that co-depend on each other across spatial and temporal scales—single cells coming together into larger and larger ensembles. These processes undergo constant rearrangements and shifts, balancing on the boundary of criticality, striving to remain far from thermodynamic equilibrium. What "Generality" the system and its parts exhibit results from constant flux, from its Heraclitean nature, where constant change is required to remain in the same place.

Hence, the task of engineering AR must not be approached from the top-down, and not only because of the high computational complexity of coarse-grained cognitive functions (discussed in Section Traditional vs. fine-grained functionalism). Soft materials need to be engineered into simple "proto-cognitive" units which then need to be scaled up into higher-level systems. While we believe that to a large degree appropriate scaling up requires self-organization, the researcher still remains largely in control of this process, as they can influence and shape the fitness landscape of emerging autonomous embodied robots, guiding them toward meta-stable states that they deem beneficial or useful. In fact, to a degree even an external re-arrangement of the emerging self-organized system may be enough to push it in a particular direction, in a manner similar to how surgical intervention into grown tissue makes possible the creation of xenobots (Kriegman et al., 2020).

This would mean that the task of developing AR doing a wide range of things—whether that would be driving cars, repairing spaceships, performing surgeries, or accompanying us at the table—is likely beyond the limits of what is attainable in the lifetime of the current generation of AI researchers. We believe, however, that the strategy remains similar whether one focuses on this kind of blue-sky research, or rather seeks to achieve more proximal goals that are already stated in the literature among the things engineers are working toward. These more feasible applications, specifically in the case of xenobots, include "intelligent drug delivery", "internal surgery", identifying cancer or processing of toxic waste products (listed by Kriegman et al., 2020), as well as "cleaning microfluidic chambers" and "environmental sensing" (suggested by Blackiston et al., 2021). The common approach to the development of such machines focuses on what conditions would be required for the system to believe this task to be "good"—not in terms of arbitrary reward functions, but in terms of risks and opportunities, or fitness landscapes. One way to accomplish this goal may be in parallel to raising and educating a child (cf. Ciaunica et al., 2021). In contrast to standard approaches in contemporary AI research, which may be more accurately compared with operant conditioning, raising a child consists more in creating—and removing—affordances in the social and physical environment of the baby. We create opportunities, control some of the risks, but in the end it is the child that must take up any particular affordance in order to best learn it. We reward, correct, and punish, but most often we do so implicitly, by accident, and to a much lesser degree than in the case of AI systems. These sparse rewards can be taken to serve more to structure the fitness landscape that the child explores, to boost its internal reward and motivation systems, than to provide a reward or loss function that learning can entirely depend upon.

The approach toward AR we suggest is similar. In the—paradoxically—simplest case where we rely on living soft materials as building blocks, we can observe an application of this strategy in the case of the aforementioned xenobots. In a virtual cyborg-like setup, they explore in simulation their expected fitness landscape guided by some simple

tasks and then, *in vivo*, the simple self-organized structure is finessed through external means. The resulting living robot is capable of surprisingly complex behavioural feats, as it forages throughout its simple environment on a Petri dish, coordinating its behaviour with others, and—when presented with an opportunity of interacting with "naive" stem cells—replicating into active organisms, similar in form and abilities (Kriegman et al., 2021). Xenobots, in fact, offer an initial hint that an approach along the lines we suggested here is feasible and may well lead to the development of workable AR, even if with only limited applications to begin with.

## Summary

In sum, transitioning toward AR capable of selecting their own goals requires incorporating "dynamic materials that possess a substantial degree of conformational freedom, mobility, and exchange of nanoscale components" (Kaspar et al., 2021: 353). In other words, developing these autonomous systems calls for attending to the fine-grained functional profile of embodied active materials that do not themselves depend on exogenous control. This is in fact one of the main outstanding problems in synthetic biology and soft robotics, as current model systems are not capable of self-organizing in a coordinated manner across the nano and mesoscales, instead relying on researchers painstakingly guiding the process to a desired state. In point of fact, how organisms themselves are able to develop toward species-invariant morphological outcomes is itself an open question of momentous importance in biology. However, there is hope that exploiting what we already know about multicellular development and the physical principles of self-organization—paying attention to appropriate "scaling up" of intelligent parts into wholes—can help make way on this in the synthetic domain.

Our suggestion, then, follows recent lines of research that emphasise the importance of constructing machines that themselves comprise smart, active, and, in some cases, cognitive parts. Sometimes this is phrased as a matter of "off-loading" computation from centralised computers to the body, though the language of embodied computation is ambiguous and difficult to specify technically [see Nakajima et al. (2015) and Müller and Hoffman (2017) for divergent stances on this]. What does seem crucial for this next stage of designed soft robots is the ability to achieve global coordination in a more autonomous and self-organized manner, something currently out of reach but hopefully not for too long, as that is itself an active area of research.

# Conclusion

We opened this paper with the suggestion that to create autonomous embodied robots capable of valuing and engaging with the world in a goal-directed manner requires incorporating several dimensions of biological endowment. In particular, we looked at the notion of precarity ("risk-to-self") and the related notion of existential needs. The proposal here has been that AI and AR come to autonomously value and interact with the world the more they approach biological analogs thereof–and the closer they approximate the dynamics that introduce the possibility of existential consequences for its actions. In other words, the proposal here has been to develop cognitive sophistication within alternative media by incorporating dimensions of vulnerability, precarity, and existential needs that emanate from the system's own internal dynamics with a denumerable set of actions that must be taken for this system to remain in far-from-equilibrium conditions.

To this end, we set out to dislodge several key assumptions embedded in the cognitive sciences that undermine the crucial role materiality plays in instantiating, mediating, and enabling cognitive form and function–specifically by proposing a fine-grained functionalist approach that treats the more minute properties of the system as central for cognitive function. The fields of active matter physics and soft robotics have begun to blur long-held dichotomies between hardware and software, living and non-living, machine and organism, and so on. But rather than reducing organisms to an anachronistic understanding of mechanism or matter, these fields have begun to actively question our understanding of materiality entirely. What we find, then, is not the hard, inert, and wholly passive parts standardly associated with machines and robots—but an inherent activity suffused throughout certain materials that, when brought into concerted, guided, and orchestrated engagement with one another *via* bioelectricity, can manifest and expand machine functionality in a manner unavailable to paradigms that do not avail themselves of these techniques. To construct autonomous robots, then, we propose an explicitly thermodynamic conception of life and mind that expands the domain of both terms. In effect, the view we have tried to articulate is one in which the mind is more material, and the material more mental, than is commonly believed—a view that is more at home in 18th and 19th century romanticist thought and American pragmatism (e.g., Charles Sanders Peirce) than it is with 20th century reductive theories of matter. When we shift our perspective away from one in which higher-level cognition sits across a divide from inert, passive matter to a view in which materiality is already pregnant with the possibility of the mental, we believe we move one step

closer to the goal of creating autonomous and hence actually intelligent machines.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## Author contributions

DH, WR, and UL have equally conceived and developed the main thesis of the paper and edited, prepared, and accepted the final draft for publication. DH and WR wrote the first draft. All authors contributed to the article and approved the submitted version.

## Funding

This paper was generously funded by the Leverhulme Trust, through the Leverhulme Centre for the Future of Intelligence.

## Acknowledgments

This paper was born as a result of the Diverse Intelligences Summer Institute in July 2021, held online, where the authors met. We would like to express our gratitude to our original project members from DISI: Alejandra Ciria, Nischal Mainali, and Dominic Sivitilli, for the many conversations and inspirations that led to the writing of this paper, and the unofficial member of the group-Taylor Beck. We would also like to thank the DISI 2021 directors-Erica Cartmill, Kensy Cooperrider, and Jacob Foster-for creating and fostering such a unique interdisciplinary environment. Finally, we would like to acknowledge the impact of discussions with and comments from Michael Levin and Marta Halina, as well as the audiences at the 17th Annual Conference of the Italian Association of Cognitive Sciences in Noto, December 2021, the 2nd International Conference on Embodied Intelligence held online in March 2022, and the British Society for the Philosophy of Science (BSPS) Annual Conference in Exeter, July 2022. UL extends gratitude to the Austrian Marshall Plan Foundation for their support.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Arias Del Angel, J. A., Nanjundiah, V., Benítez, M., and Newman, S. A. (2020). Interplay of mesoscale physics and agent-like behaviors in the parallel evolution of aggregative multicellularity. *EvoDevo* 11, 1–18. doi: 10.1186/s13227-020-00165-8

Baez, J. C. (2021). The Joy of Condensed Matter. Nautilus Magazine. Available online at: https://nautil.us/the-joy-of-condensed-matter-9608/ (accessed August 11, 2022).

Baluška, F., and Levin, M. (2016). On having no head: Cognition throughout biological systems. *Front. Psychol.* 7, 902. doi: 10.3389/fpsyg.2016.00902

Batterman, R. (2021). "Multiscale modelling in inactive and active materials," in *Levels of Biological Organization in the Biological Sciences.* eds, D. Brooks, J. DiFrisco, and W. C. Wimsatt (Cambridge, Massachusetts: MIT Press).

Bechtel, W., and Bich, L. (2021). Grounding cognition: heterarchical control mechanisms in biology. *Philos. Transac. R. Soc. B.* 376, 20190751. doi: 10.1098/rstb.2019.0751

Bickhard, M. (1993). Representational content in humans and machines. *J. Exper. Theor. Artif. Intell.* 5, 285–233. doi: 10.1080/09528139308953775

Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., and Levin, M. (2021). A cellular platform for the development of synthetic living machines. *Sci. Robot.* 6, eabf1571. doi: 10.1126/scirobotics.abf1571

Block, N. (1978). Troubles with functionalism. *Minnesota Stud. Philos. Sci.* 9, 261–325.

Bongard, J., and Levin, M. (2021). Living things are not (20th century) machines: Updating mechanism metaphors in light of the modern science of machine behavior. *Front. Ecol. Evol.* 9, 650726. doi: 10.3389/fevo.2021.650726

Boussard, A., Fessel, A., Oettmeier, C., Briard, L., Döbereiner, H. G., and Dussutour, A. (2021). Adaptive behaviour and learning in slime moulds: the role of oscillations. *Philos. Transac. R. Soc. B.* 376, 20190757. doi: 10.1098/rstb.2019.0757

Brugues, J., and Needleman, D. (2014). Physical basis of spindel self-organization. *Proc. Nat. Acad. Sci.* 111, 18496–18500. doi: 10.1073/pnas.1409404111

Brunet, T. D. P., and Halina, M. (2020). Minds, machines, and molecules. *Philos. Topic.* 48, 221–241. doi: 10.5840/philtopics202048111

Cejkova, J., Novak, M., Stepanek, F., and Hanczyc, M. (2014). Dynamics of chemotactic droplets in salt concentration gradients. *Langmuir* 30, 11937–11944. doi: 10.1021/la502624f

Chirimuuta, M. (2018). Marr, Mayr, and MR: What functionalism should now be about. *Philos. Psychol.* 31, 403–418. doi: 10.1080/09515089.2017.1381679

Ciaunica, A., Safron, A., and Delafield-Butt, J. (2021). Back to square one: the bodily roots of conscious experiences in early life. *Neurosci. Conscious.* 2021, niab037. doi: 10.31234/osf.io/zspm2

Cichos, F., Gustavsson, K., Mehlig, B., and Volpe, G. (2020). Machine learning for active matter. *Nat. Mach. Intell.* 2, 94–103. doi: 10.1038/s42256-020-0146-9

Dalous, J., Burghardt, E., Müller-Taubenberger, A., Bruckert, F., Gerisch, G., and Bretschneider, T. (2008). Reversal of cell polarity and actin-myosin cytoskeleton reorganization under mechanical and chemical stimulation. *Biophys. J.* 94, 1063–1074. doi: 10.1529/biophysj.107.114702

Di Paolo, E., Buhrmann, T., and Barandiaran, X. E. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198786849.001.0001

Dinet, C., Michelot, A., Herrou, J., and Mignot, T. (2021). Linking single-cell decisions to collective behaviours in social bacteria. *Philos. Transac. R. Soc. B* 376, 20190755. doi: 10.1098/rstb.2019.0755

Ebrahimkhani, M. R., and Levin, M. (2021). Synthetic living machines: a new window on life. *IScience* 24, 102505. doi: 10.1016/j.isci.2021.102505

Egbert, M. (2022). *Towards Adaptive Sensorimotor Autonomy: Developing A System That can Adapt to Its Own Emergent and Dynamic Needs*. doi: 10.1162/isal_a_00503

Egbert, M. D. (2021). Self-preserving Mechanisms in Motile Oil Droplets: A computational model of abiological self-preservation. *R. Soc. Open Sci.* 8, 210534. doi: 10.1098/rsos.210534

Fields, C., Bischof, J., and Levin, M. (2020). Morphological coordination: a common ancestral function unifying neural and non-neural signalling. *Physiology* 35, 16–30. doi: 10.1152/physiol.00027.2019

Fields, C., Glazebrook, J. F., and Levin, M. (2021). Minimal physicalism as a scale-free substrate for cognition and consciousness. *Neurosci. Consc.* 2021, niab013. doi: 10.1093/nc/niab013

Fields, C., and Levin, M. (2020). Scale-free biology: integrating evolutionary and developmental thinking. *BioEssays* 42, 1900228. doi: 10.1002/bies.201900228

Fodor, E., and Marchetti, C. (2018). The statistical physics of active matter: From self-catalytic colloids to living cells. *Phys. A Stat. Mech. Appl.* 504, 106–120. doi: 10.1016/j.physa.2017.12.137

Friston, K. J. (2019). A free energy principle for a particular physics. ArXiv E-Prints, arXiv:1906.10184.

Froese, T. (2016). "Life is precious because it is precarious: individuality, mortality, and the problem of meaning," in *Representation of Reality: Humans, Other Living Organisms and Intelligent Machines*, eds. D. Dodig-Crnkovic, and R. Giovagnoli (Springer: Berlin). doi: 10.1007/978-3-319-43784-2_3

Gershman, S. J., Balbi, P. E., Gallistel, C. R., and Gunawardena, J. (2021). Reconsidering the evidence for learning in single cells. *ELife*. 10, e61907. doi: 10.7554/eLife.61907

Godfrey-Smith, P. (2016a). Mind, matter, and metabolism. *J. Philos*. 113, 481–506. doi: 10.5840/jphil20161131034

Godfrey-Smith, P. (2016b). Individuality, subjectivity, and minimal cognition. *Biol. Philos.* 3, 275–286. doi: 10.1007/s10539-016-9543-1

Godfrey-Smith, P. (2017). The subject as cause and effect of evolution. *Interface Focus* 7, 22. doi: 10.1098/rsfs.2017.0022

Goertzel, B., and Pennachin, C. (2007). *Artificial General Intelligence*. Cognitive Technologies. Berlin: Springer. doi: 10.1007/978-3-540-68677-4

Hanczyc, M. (2011). Metabolism and motility in prebiotic structures. *Philos. Trans. R. Soc. B* 366, 2885–2893. doi: 10.1098/rstb.2011.0141

Hanczyc, M., and Ikegami, T. (2010). Chemical basis for minimal cognition. *Artif. Life* 16, 233–243. doi: 10.1162/artl_a_00002

Hanson, A. (2021). Spontaneous electrical low-frequency oscillations: a possible role in *hydra* and all living systems. *Philos. Transac. R. Soc. B.* 376, 20190763. doi: 10.1098/rstb.2019.0763

Haugeland, J. (1998). *Understanding Natural Language*. Harvard University Press.

Hodges, B., and Baron, R. (1992). Values as constraints on affordances: perceiving and acting properly. *J. Theory Soc. Behav.* 22, 263–294. doi: 10.1111/j.1468-5914.1992.tb00220.x

Hodges, B., and Raczaszek-Leonardi, J. (2021). Ecological values theory: beyond conformity, goal-seeking, and rule-following in action and interaction. *Rev. General Psychol.* 26, 86–103. doi: 10.1177/10892680211048174

Jekely, G. (2021). The chemical brain hypothesis for the origin of nervous systems. *Philos. Transac. R. Soc. B* 376, 761. doi: 10.1098/rstb.2019.0761

Jékely, G., Keijzer, F., and Godfrey-Smith, P. (2015). An option space for early neural evolution. *Philos. Transac. R. Soc. B.* 370, 20150181. doi: 10.1098/rstb.2015.0181

Jones, J. (2015). Applications of multi-agent slime mould computing. *Int. J. Parallel, Emerg. Distr. Syst.* 31, 420–449.

Juarrero, A. (2015). What does the closure of context-sensitive constraints mean for determinism, autonomy, self-determination, and agency? *Progress in Biophys. Molec. Biol.* 119, 510–521. doi: 10.1016/j.pbiomolbio.2015.08.007

Jülicher, F., Grill, S. W., and Salbreux, G. (2018). Hydrodynamic theory of active matter. *Reports Progr. Phys.* 81, 076601. doi: 10.1088/1361-6633/aab6bb

Kanao, K., Harada, S., Yamamoto, Y., Honda, W., Aerie, T., Akita, S., et al. (2015). Highly selective flexible tactile strain and temperature sensors against substrate bending for an artificial skin. *RSC Adv.* 5, 30170. doi: 10.1039/C5RA03110A

Kaspar, C., Ravoo, B. J., van der Wiel, W. G., Wegner, S. V., and Pernice, W. H. P. (2021). The rise of intelligent matter. *Nature* 594, 345–355. doi: 10.1038/s41586-021-03453-y

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press. doi: 10.1007/978-94-015-8054-0_8

Keijzer, F., Van Duijn, M., and Lyon, P. (2013). What nervous systems do: early evolution, input–output, and the skin brain thesis. *Adapt. Behav.* 21, 67–85. doi: 10.1177/1059712312465330

Kelso, S. (2016). On the self-organising origins of agency. *Trends Cogn. Sci.* 20, 490–499. doi: 10.1016/j.tics.2016.04.004

Kempf, F., Mueller, R., Frey, E., Yeomans, J. M., and Doostmohammadi, A. (2019). Active matter invasion. *Soft Matter* 15, 7538–7546. doi: 10.1039/C9SM01210A

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Nat. Acad. Sci.* 117, 1853–1859. doi: 10.1073/pnas.1910837117

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2021). Kinematic self-replication in reconfigurable organisms. *Proc. Nat. Acad. Sci.* 118, e2112672118. doi: 10.1073/pnas.2112672118

Levin, M. (2019). The computational boundary of a 'Self': Developmental bioelectricity drives multicellularity and scale-free cognition. *Front. Psychol.* 10, 2688. doi: 10.3389/fpsyg.2019.02688

Levin, M. (2020). The biophysics of regenerative repair suggests new perspectives on biological causation. *BioEssays.* 42, 1900146. doi: 10.1002/bies.201900146

Levin, M. (2021). Life, death, and self: Fundamental questions of primitive cognition viewed through the lens of body plasticity and synthetic organisms. *Biochem. Biophys. Res. Commun.* 564, 114–133. doi: 10.1016/j.bbrc.2020.10.077

Levin, M., and Dennett, D. C. (2020). *Cognition All The Way Down*. Aeon. Available online at: https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas (accessed November 8, 2022).

Lyon, P. (2006). The biogenic approach to cognition. *Cogn. Process.* 7, 11–29. doi: 10.1007/s10339-005-0016-8

Lyon, P. (2015). The cognitive cell: Bacterial behavior reconsidered. *Front. Microbiol.* 6, 264. doi: 10.3389/fmicb.2015.00264

Lyon, P., Keijzer, F., Arendt, D., and Levin, M. (2021). Reframing cognition: getting down to biological basics. *Philos. Transac. R. Soc. B* 376, 750. doi: 10.1098/rstb.2019.0750

Lyon, P., and Kuchling, F. (2021). Valuing what happens: A biogenic approach to valence and (potentially) affect. *Philos. Trans. Royal Soc. B. Biol. Sci.* 376, 20190752. doi: 10.1098/rstb.2019.0752

Man, K., and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* 1, 446–452. doi: 10.1038/s42256-019-0103-7

Marchetti, M. C., Joanny, J. F., Ramawamy, S., Liverpool, T. B., Prost, J., Madan, R., et al. (2013). Hydrodynamics of sft active matter. *Rev. Modern Phys.* 85, 1143–1189. doi: 10.1103/RevModPhys.85.1143

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living, Vol. 42*. Springer Netherlands. doi: 10.1007/978-94-009-8947-4

McCarthy, J., and Hayes, P. J. (1969). "Some philosophical problems from the standpoint of artificial intelligence," in *Machine Intelligence, Vol. 4*, eds B. Meltzer and D. Michie (Edinburgh: Edinburgh University Press), 463–502.

McGivern, P. (2020). Active materials: minimal models of cognition? *Adapt. Behav.* 28, 441–451. doi: 10.1177/1059712319891742

Meyer, R., and Brancazio, N. (2021). Putting down the revolt: Enactivism as a philosophy of nature. *Front. Psychol.* 13, 948733. doi: 10.3389/fpsyg.2022.948733

Mogilner, A., and Manhart, A. (2018). Intracellular fluid mechanics: Coupling cytoplasmic flow with active cytoskeletal gel. *Ann. Rev. Fluid Mech.* 50, 347–370. doi: 10.1146/annurev-fluid-010816-060238

Moreno, A., and Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. London: Springer.

Müller, V., and Hoffman, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artif. Life* 23, 1–24. doi: 10.1162/ARTL_a_00219

Nakajima, K., Hauser, H., Li, T., and Pfeifer, R. (2015). Information processing via physical soft body. *Sci. Rep.* 5, 1–11. doi: 10.1038/srep10487

Needleman, D., and Dogic, Z. (2017). Active matter at the interface between materials science and cell biology. *Nat. Rev. Mater.* 2, 1–14. doi: 10.1038/natrevmats.2017.48

Newman, S. (2021). "Inherency," in *Evolutionary Developmental Biology: A Reference Guide, Volume 1*, eds L. N. de la Rosa and G. Muller (Cham: Springer), 121–132. doi: 10.1007/978-3-319-32979-6_78

Newman, S. (2022). Inherency and agency in the origin and evolution of biological function. *Biol. J. Linnean Soc.* 20, 1–16. doi: 10.1093/biolinnean/blac109

Newman, S. A. (2016). "Biogeneric'developmental processes: drivers of major transitions in animal evolution. *Philos. Transac. R. Soc. B.* 371, 20150443. doi: 10.1098/rstb.2015.0443

Newman, S. A. (2019). Inherency of form and function in animal development and evolution. *Front. Physiol.* 10, 702. doi: 10.3389/fphys.2019.00702

Newman, S. A. (2020). "The origins and evolution of animal identity," in *Biological Identity: Perspectives from Metaphysics and the Philosophy of Biology,* eds. A. S. Meincke and J. Dupre (London: Routledge). doi: 10.4324/9781351066389-7

Nicholson, D. (2014). The machine conception of the organism in development and evolution: A critical analysis. *Stud. History Philos. Biol. Biomed. Sci.* 48, 162–174. doi: 10.1016/j.shpsc.2014.08.003

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progr. Neurobiol.* 134, 17–35. doi: 10.1016/j.pneurobio.2015.09.001

Pfeifer, R., Lungarella, M., and Iida, F. (2007). Self-organization, embodiment, and biologically inspired robots. *Science* 318, 1088–1093. doi: 10.1126/science.1145803

Pfeifer, R., and Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press.

Pishvar, M., and Harne, R. L. (2020). Foundations for soft, smart matter by active mechanical metamaterials. *Adv. Mater.* 7, 2001384. doi: 10.1002/advs.202001384

Polger, T. W., and Shapiro, L. A. (2016). *The Multiple Realization Book, 1st Edn*. Oxford University Press.

Poon, W. C. K. (2000). "A day in the life of a hard-sphere suspension," in *Soft and Fragile Matter: Nonequilibrium Dynamics, Metastability, and Flow.* eds, M. E. Cates, and M. R. Evans. (Bristol; Philadelphia: Edinburgh, Scotland: Institute of Physics Pub.; copublished by Scottish Universities Summer School in Physics). doi: 10.1201/9781420033519.ch1

Popkin, G. (2016). The physics of life. *Nat. News* 529, 16–18. doi: 10.1038/529016a

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: The University of Chicago Press. doi: 10.7208/chicago/9780226507194.001.0001

Potochnik, A. (2021). *Our World Isn't Organised Into Levels*. In Daniel Brooks, James DiFrisco, and William Wimsatt Levels of Organization in the Biological Sciences. Cambridge, Mass: MIT Press.

Prindle, A., Liu, J., Asally, M., Ly, S., Garcia-Ojalvo, J., and Süel, G. M. (2015). Ion channels enable electrical communication in bacterial communities. *Nature* 527, 59–63. doi: 10.1038/nature15709

Putnam, H. (1975). *Mind, Language, and Reality: Philosophical Papers*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511625251

Putnam, H. (1988). *Representation and Reality*. Cambridge, MA: MIT Press.

Reynolds, A. (2007). The theory of the cell state and the question of cell autonomy in nineteenth and early twentieth-century biology. *Sci. Context* 20, 71–95. doi: 10.1017/S0269889706001141

Rorot, W. (2022). Counting with cilia: The role of morphological computation in basal cognition research. *Entropy,* 24, 1581. doi: 10.3390/e24111581

Sanchez, T., Chen, D. T., DeCamp, S. J., Heymann, M., and Dogic, Z. (2012). Spontaneous motion in hierarchically assembled active matter. *Nature* 491, 431–434. doi: 10.1038/nature11591

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588, 604–609. doi: 10.1038/s41586-020-03051-4

Shah, D. S., Powers, J. P., Tilton, L. G., Kriegman, S., Bongard, J., and Kramer-Bottiglio, R. (2021). A soft robot that adapts to environments through shape change. *Nat. Mach. Intell.* 3, 51–59. doi: 10.1038/s42256-020-00263-1

Sims, M. (2020). How to count biological minds: symbiosis, the free energy principle, and reciprocal multiscale integration. *Synthese* 199, 2157–2179. doi: 10.1007/s11229-020-02876-w

Sims, M. (2021). A continuum of intentionality: linking the biogenic and anthropogenic approaches to cognition. *Biol. Philos.* 38, 1–31. doi: 10.1007/s10539-021-09827-w

Smith-Ferguson, J., and Beekman, M. (2020). Who needs a brain? Slime moulds, behavioural ecology and minimal cognition. *Adapt. Behav.* 28, 465–478. doi: 10.1177/1059712319826537

Thompson, E. (2010). *Mind in Life: Biology, Phenomenology, and The Sciences Of Mind*. Harvard University Press.

Van Duijn, M., Keijzer, F., and Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adapt. Behav.* 14, 157–170. doi: 10.1177/105971230601400207

Varela, F., Rosch, E., and Thompson, E. (1993). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, Massachusetts: MIT Press.

Wan, K. Y., and Jekely, G. (2021). Origins of eukaryotic excitability. *Philos. Transac. R. Soc. B* 376: 20190758.20190758. doi: 10.1098/rstb.2019.0758

Weber, A., and Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenol. Cogn. Sci.* 1, 97–125. doi: 10.1023/A:1020368120174

Wioland, H., Lushi, E., and Goldstein, R. E. (2016). Directed collective motion of bacteria under channel confinement. *New J. Phys.* 18, 075002. doi: 10.1088/1367-2630/18/7/075002

Wu, C., Dai, J., Li, X., Gao, L., Wang, J., Liu, J., et al. (2021). Ion-exchange enabled synthetic swarm. *Nat. Nanotechnol.* 16, 288–295. doi: 10.1038/s41565-020-00825-9

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 21 December 2022
DOI 10.3389/fnbot.2022.847054

Check for updates

OPEN ACCESS

EDITED BY
Inês Hipólito,
Humboldt University of Berlin,
Germany

REVIEWED BY
Alban Laflaquière,
Softbank Robotics, France
David Vernon,
Carnegie Mellon University Africa,
Rwanda

*CORRESPONDENCE
Matthew D. Egbert
✉ m.egbert@auckland.ac.nz

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Neurorobotics

RECEIVED 01 January 2022
ACCEPTED 29 September 2022
PUBLISHED 21 December 2022

CITATION
Egbert MD and Barandiaran XE (2022) Using enactive robotics to think outside of the problem-solving box: How sensorimotor contingencies constrain the forms of emergent autononomous habits.
*Front. Neurorobot.* 16:847054.
doi: 10.3389/fnbot.2022.847054

COPYRIGHT
© 2022 Egbert and Barandiaran. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Using enactive robotics to think outside of the problem-solving box: How sensorimotor contingencies constrain the forms of emergent autononomous habits

Matthew D. Egbert[1,2]* and Xabier E. Barandiaran[3]

[1]School of Computer Science, University of Auckland, Auckland, New Zealand, [2]Te Ao Mārama, University of Auckland, Auckland, New Zealand, [3]IAS-Research Center for Life, Mind, and Society, Department of Philosophy, University of the Basque Country (UPV/EHU), Bilbao, Spain

We suggest that the influence of biology in 'biologically inspired robotics' can be embraced at a deeper level than is typical, if we adopt an enactive approach that moves the focus of interest from how problems are solved to how problems emerge in the first place. In addition to being inspired by mechanisms found in natural systems or by evolutionary design principles directed at solving problems posited by the environment, we can take inspiration from the precarious, self-maintaining organization of living systems to investigate forms of cognition that are also precarious and self-maintaining and that thus also, like life, have their own problems that must be be addressed if they are to persist. In this vein, we use a simulation to explore precarious, self-reinforcing sensorimotor habits as a building block for a robot's behavior. Our simulations of simple robots controlled by an Iterative Deformable Sensorimotor Medium demonstrate the spontaneous emergence of different habits, their re-enactment and the organization of an ecology of habits within each agent. The form of the emergent habits is constrained by the sensory modality of the robot such that habits formed under one modality (vision) are more similar to each other than they are to habits formed under another (audition). We discuss these results in the wider context of: (a) enactive approaches to life and mind, (b) sensorimotor contingency theory, (c) adaptationist vs. structuralist explanations in biology, and (d) the limits of functionalist problem-solving approaches to (artificial) intelligence.

KEYWORDS

**enactive robotics, sensorimotor contingencies, problem-solving, habits, IDSM, sensorimotor autonomy**

# 1. Introduction

Artificial Intelligence and the scientific approach to mind (what is known as *cognitive science*) was born (or rather raised) as a problem solving discipline (Newell et al., 1958; Putnam, 1965; Fodor, 1968). Deprived of life, the machine metaphor was one of symbol manipulation and rationality (deductive, inferential, heuristic or otherwise). The unprecedented potential of Universal Turing Machines (computers) was the driving metaphor to study the mind. The software, the mind, was the problem solving method, the hardware, the brain, its implementation.

Alternative conceptions of the mind were available at the origins of Artificial Intelligence and Robotics (Grey Walter, 1950; Ashby, 1952) but the rapid success of computer science left them aside. Over time, the limitations of the problem-solving centered computational theory of the mind became apparent and the biologically inspired, embodied and later enactive conceptions of the mind gained momentum. We are now immersed in a mesh of hybrid architectures, applied to a wide range of practices, from industrial to scientific modeling applications, and a new summer of Artificial Intelligence is rising, with robotics as a major container of social and technological expectations.

There are good reasons for why problem-solving attracts so much attention from researchers, but it is pertinent to ask: what aspects of minds are omitted or obscured by the problem-solving focused perspective? what can life teach us about what intelligence holds before and beyond problem-solving? and even when problem solving is addressed …how is it that natural agents have and become concerned by *their own* problems?

This paper has two goals. The first is to argue that by abandoning problem solving (or at least putting it down for a time), other useful explanatory targets and ways to explain minds are given space to emerge.

The second, more specific goal is a case in point: we use simulated robots to show how sensorimotor contingencies influence the formation of self-maintaining patterns of sensorimotor activity "habits" in regular ways that depend upon sensory modality. By de-emphasizing problem solving, we are able to take a fresh look at the relationship between sensory modalities, sensorimotor contingencies and habitual behaviors. But to explain these results, we first need to provide more context.

The paper proceeds as follows. The next section explains what we mean by "problem-solving," why it has been a popular target within the cognitive sciences, and what we see as the primary disadvantage of excessive attention being given to the topic. We then explore the intimate relationship between robotics and biologically inspired and embodied problem-solving paradigms. Section 2.3 introduces the enactivist concept of autonomy, providing an alternative framework for developing Sensorimotor Contingency Theory outside of the problem-solving approach. The remainder of the paper presents and analyzes a simulation model that is used to explain: (i) that robots must first have their own problems instead of solving those posited by external observers; (ii) that, in doing so, they must assert a way of life whose structure and form must be taken as the object of study. We finally discuss some of the larger implications of our enactive approach in connection with wider theories of biological explanation and inspiration.

# 2. From problem solving to enactive robot

## 2.1. Problem-solving in minds and machines

We use the term "problem-solving" to refer, in a broad and inclusive manner, to the kinds of things that we associate with being capable or clever. Nowadays, for many, "the ability to solve problems is not just an aspect or feature of intelligence—it is the essence of intelligence" (Hambrick et al., 2020, p.553). It is certainly not a new idea. The very birth of Artificial Intelligence owes much to it (Newell et al., 1958). As Newell and Simon later stated: "Since ability to solve problems is generally taken as a prime indicator that a system has intelligence, it is natural that much of the history of artificial intelligence is taken up with attempts to build and understand problem-solving systems" (Newell and Simon, 1976, p. 120). The task of artificial intelligence was thus to devise potential solution-generators and to design tests that could evaluate them. This assumed that the problem space was well fixed so that solutions could both be evaluated and generated. Decades later, Artificial Intelligence handbooks still devote their first central sections to problem solving (e.g., see part II of Norvig and Russell, 2021, 4th edition).

A problem is a context in which behaviors can be evaluated according to a norm of success at "solving" the problem. Problems vary from being trivial to challenging to impossible as the proportion of behaviors that are good (out of all possible behaviors) shrinks. They include "high-level" human problem-solving, such as the skills that are taught in schools or universities, as well as embodied problems such as balancing on two feet or swimming efficiently. For a system to be evaluated in terms of problem-solving, one must first have the specification of the context and of a normative evaluation so that behaviors within that context can be compared as more or less successful. Defined as such, just about anything can be seen as a problem solver. A bottle lid solves the problem of preventing spills; a car's differential solves the problem of distributing force effectively to its wheels; a computer program solves the problem of beating a human at chess.

Herein lies both the advantage and the disadvantage of placing problem solving at the center of the cognitive sciences: *almost anything can be evaluated in terms of its problem solving*

*ability*. On the one hand this is a great boon. Quantifying problem solving ability is relatively straight-forward and the ability to quantify how effective a system is at solving one or more problems facilitates technoscientific progress where measurements can play an important role in defining progress. This is apparent in artificial intelligence (AI) research where benchmarks such as chess and other games (Canaan et al., 2019), hand-writing recognition (e.g., Graves and Schmidhuber, 2009), image classification, speech recognition, etc. (e.g., MLPerf benchmarks MLPerf, 2021), are used to quantify progress. Problem-solving similarly provides metrics for studying the minds in psychology and neuroscience, where problem-solving related notion of 'tasks' (e.g., the Simon task) are used to structure human activity and performance related metrics such as reaction speed or error rate are seen as providing key insights into how our minds operate (Simon and Wolf, 1963).

The ease of measuring problem-solving ability sometimes leads to it (problem-solving ability) becoming the *explanandum*—the thing we strive to understand. This is seen in research questions like: *How do people recognize faces so well? How do babies come to understand the motives of other people? How do we play chess? How can we make a safe self-driving car?* etc. Problem-solving also sometimes becomes the *explananda*—the terms in which we explain what minds are, how they work, or what they do. Evolutionary psychology (Pinker, 1997; Buss, 1998), for example, emphasizes the evolutionary advantage of problem-solving ability and in this context, explanations that are provided in terms of problem-solving ability are seen as complete, as evolution can be invoked to explain why such mechanisms exist. The evolutionary advantage of having a mind is in its contribution to problem solving and therefore, minds are best understood as problem solving machines.

However, it has been argued that: "The essence of intelligence is to act appropriately when there is no simple pre-definition of the problem or the space of states in which to search for a solution. Rational search within a problem space is not possible until the space itself has been created, and is useful only to the extent that the formal structure corresponds effectively to the situation" (Winograd and Flores, 1987, p.98). Thus, even from a problem solving perspective, intelligence is not really the capacity to solve a problem but to bring a situation into a fabricated frame where it can be treated as a problem to be solved.

Moreover, the problem with excessive focus upon problem-solving is that there are other unique and important features of minds that are worthy of study—features that may only indirectly relate to problem-solving ability or perhaps not at all. The problem, in a nutshell, is the conflation of (i) "problem-solving ability" with (ii) all of the other phenomena associated with "being a mindful body."

The mainstream computational functionalist approach to the mind (Putnam, 1965; Fodor, 1968) doesn't really help much addressing what mindful bodies are beyond problem-solving devices. For Putnam, the very definition of the mental is always in reference to a Turing machine table that works out rational transitions (e.g., computing and storing preferences over a utility function or solving problems in problem representation space). Deviations from this rationality are treated as pathological. All human mental life is, according to Putnam, not perfectly normal, thus relatively pathological. Putnam acknowledges "our model is an overly simple and overly rationalistic one in a number of respects. However, it would be easy, in principle, although perhaps impossible in practice, to complicate our model in all these respects—to make the model dynamical, to allow for irrationalities in preference, to allow for irrationalities in the inductive logic of the machine, to allow for deviations from the rule: maximize the estimated utility. But I do not believe that any of these complications would affect the philosophical conclusions reached in this paper" (Putnam, 1965, p. 43).

Deviations from the abstract rational rule are pathological. Explanation lies on the pure domain of abstract problem solving, the deviations make it all more complicated (as if dirt in the form of a set of exceptions where to be added to the pure explanation) but change fundamentally nothing. As we are about to see, embodied approaches, and particularly enactivism, bring these "pathological" expressions to the center of the explanation (of which rational thinking is the complicated achievement) turning it into the core constitution of mind. Biologically inspired robotics has a lot to contribute in this direction.

## 2.2. From biologically inspired problem-solving to enactive robotics

Enactivism was born under the conviction that robotics, as a field, would require, or even force, cognitive science to move beyond the problem-solving framework:

> The assumption in CS [Cognitive Science] has all along been that the world can be divided into regions of discrete elements and tasks to which the cognitive system addresses itself, acting within a given "domain" of problems: vision, language, movement. Although it is relatively easy to define all possible states in the "domain" of the game of chess, it has proven less productive to carry this approach over into, say, the "domain" of mobile robots. Of course, here too one can single out discrete items (such as steel frames, wheels and windows in a car assembly). But it is also clear that while the chess world ends neatly at some point, the world of movement amongst objects does not. It requires our continuous use of common sense to configure our world of objects. (Varela, 1992, p.251)

Yet, moving away from the problem-solving paradigm has taken a long path, most of which, rather than abandoning problem-solving has deeply transformed the way we understand how nature solves those problems. In a sense, biologically inspired robotics has mostly followed the problem-solving approach and biological inspiration has focused on picking up biological mechanisms to solve problems: from the internal neuronal inspiration of artificial neural networks since its early conception (Rosenblatt, 1958) to their later development (Rumelhart et al., 1988) to the embodied strategies that either transform the problems to be solved by their "brains" or have outsourced the computational load of the problem solving to body and world (Pfeifer and Scheier, 2001). What radically distinguished biologically inspired robotics from GOFAI (Good Old Fashioned Artificial Intelligence) was a change of focus from the abstract to the concrete, from the symbolic to the sensorimotor and from the rational to the practical know-how of situated action. Despite the emphasis on self-organization, agent-environment emergence of behavioral functioning (Steels, 1990) embodiment and situated action (Maes, 1990), etc. the main goal was still to build robots capable of solving behavioral problems. After all, to put it with biologically-inspired roboticist Barbara Webb: "The sensorimotor problems faced by animals and by robots have much in common" (Webb, 1995, p. 117) and, not only can animals help us devise robots that solve problems in a biologically inspired manner, but also solving a sensorimotor problem with the robot could help us understand how the animal solves it; like "[d]etecting which ear is closer to the sound" which "is a non-trivial problem for the cricket" (Webb, 1995, p. 120).

Other trends of biological inspiration have built upon evolutionary theory itself and artificially evolved brains or brains and bodies to solve the problems (encoded as fitness function) in what is commonly known as evolutionary robotics (Cliff et al., 1993; Nolfi et al., 2016). Random variations to the parameters of robotic brain's and bodies are selected against a fitness function that operates as the benchmark of the problem to be solved. Despite the problem-solving focus that is almost inherent in artificial evolutionary optimization techniques, evolutionary robotics served to disclose a number of principles of behavioral self-organization that non-linear, fine grained agent-environment coupling display when artificial evolution can freely explore the solution space without the prejudices inherent to the human design (Harvey et al., 1997).

Some of these approaches entail radical departures from core assumptions of the computational functionalist theory of the mind: cognitive processing does not only occur in the head and the body must be integrated as a key feature of cognitive problem solving (not simply as an executioner of the solution representation worked out in the head or a sensory transmitter of the problem into it); agent-environment interactions can self-organize with little if any representations; material bodily and interactive constitution are not mere implementation details of abstract capacities but intrinsic part of the problems and solutions that cut across them.

But *enactive robotics* moves yet further into biological inspiration. On the one hand there are enactive approaches that have attempted to introduce more of the living body of natural intelligence into robotics by including self-organized mechanical, soft bodies (Man and Damasio, 2019) or even chemical bodies (Damiano and Stano, 2021). But also, and perhaps most relevant for enactive theory transferring to the robot what metabolism has to offer to anchor intrinsic needs or emotional feedback (Ziemke and Lowe, 2009). There is however another enactive path that brings into robotics some principles of living organization and, more specifically, the autonomy of behavior, a *way of life* for robots, not as somehow transferred from the biological body, but enacted at the scale of brain-body-world dynamics (Barandiaran and Moreno, 2006). Some forerunners of this inspiration are no doubt Ross Ashby on the organism-centered inspiration of adaptive controllers as machines capable to remain homeostatic in the face of perturbations (Ashby, 1952) and Grey Walter's "life imitating" robots "designed to illustrate the uncertainty, randomness, free will or independence so strikingly absent in most well-designed machines" (Grey Walter, 1950, p. 44).

More recent development of this line of inspiration on natural and biological principles for the design of robots came hand in hand with the development of a theory of autonomous behavior and agency (Smithers, 1997), organismically inspired robotics (Di Paolo, 2003), and habit-centered enactive robotics (Egbert and Barandiaran, 2014).

## 2.3. Enactivism and the autonomy of sensorimotor life

Varela, Thompson and Rosch opened up a new way of thinking in 1991. Their enactive approach conceives that "cognitive structures emerge from the recurrent sensorimotor patterns that enable action to be perceptually guided" (Varela et al., 1991, p.173). They later stated that: "[C]ognition is no longer seen as problem solving on the basis of representations; instead, cognition in its most encompassing sense consists in the enactment or a bringing forth of a world by a viable history of structural coupling" (Varela et al., 1991, p.205). The enactive approach thus emphasizes sensorimotor coupling and the recurrent patterns that emerge from agent-environment interactions.

This way Varela overcame the operational (en)closure of the nervous system that served as his main analogy with the organization of the living, captured (together with Humberto Maturana) within the theory of autopoiesis. Ever since, the relationship between the self-organizing nature of nervous

activity and that of behavior became to some extent problematic (see Barandiaran, 2017 for a discussion).

Inspired by Maturana and Varela, Tim Smithers re-states the need for biologically inspired autonomy in robotics, in the context of the impossibility to design robot agents from a problem-solving stand point:

> Designing and building autonomous agents thus becomes the problem of designing and building processes that can support and maintain this kind of identity formation through interaction: processes that, through interaction, are continuously forming the laws of interaction that can sustain and maintain the interaction needed to form them. In other words, we need interaction processes that can support the self-construction and maintenance of interaction processes through interaction, in essentially the same way that the material and energy interaction processes of single cells can be understood as being involved in the continual forming of the mechanisms that support this interaction. Such systems will thus be self-law making as well as self-regulating, in essentially the same way as we can understand biological systems and autonomous city states. (Smithers, 1997, p.102)

This analogy between metabolic autonomy and the autonomy of behavior was further explored in Di Paolo (2003). According to this view, enactivism needs not be committed to build bio-chemically living robots (provided that this is possible or even desirable) but to endow a robot with a *way of life*. This intuition was further explored in Barandiaran (2007, 2008).

Sensorimotor Contingency Theory can further enrich this approach. It uses regularities in the ways that motor activity affect sensory activity (sensorimotor contingencies) to explain the qualities of perceptual experience (O'Regan and Noë, 2001). Empirical research involving sensory substitution, sensory modification, psychophysics has informed the development of sensorimotor contingency theory (SMCT), which attempts to explain diverse aspects of perceptual experience, including why certain sensory modalities have a particular "feel" to them; how it is possible to make one experience one sensory modality (e.g., touch) in a way that feels more like another (e.g., sight); and the conditions in which subjects are (or are not) capable of adapting to major transformations to their sensorium. The key idea in SMCT is the role that action plays in perception: a classical enactive theme (Noë, 2004).

Using variations of the basic theme of how motor activity modifies sensory input, a set of robotic architectures where made using sensorimotor contingencies as building blocks for robotic design (Maye and Engel, 2013; Jacquey et al., 2019). But these are hardly enactive in the sense of the deep biological inspiration that the enactive approach can offer.

Perhaps the best way to explore such potential is to bring forth the concept of sensorimotor autonomy (updated and refined from a previous proposal of Mental Life and also explored in more detail on the concept of Sensorimotor Life): the capacity of an agent to sustain and regulate the structures that generate behavior. This definition echoes the metabolism-based definitions of life as far-from-thermodynamic equilibrium chemical systems capable of maintaining the network of chemical reactions that constitute it (Gánti, 1975; Maturana and Varela, 1980; Rosen, 1991; Ruiz-Mirazo et al., 2004; Luisi, 2006).

The basic constituent of sensorimotor autonomy is a sensorimotor structure (a behavioral scheme or habit) made possible by both a set embodied-neural pathways and a set of sensorimotor contingency relationships. Think of it as a coordination pattern that emerges out of environmental, sensorimotor and neural (or behavior generating) mechanisms. Now, if this structure is far-from-equilibrium or, said differently, if let alone it tends to extinguish or vanish, and if the very enactment of the sensorimotor scheme reinforces itself by repetition or by satisfying certain conditions that feed-back into its supporting structure (e.g., reinforcement of synaptic connections by Hebbian learning or reward reinforcement), then we have first sense of self-maintenance that is characteristic of habits. The more the habit is enacted the more it is strenghthened, the stronger it is the more likely it is to be enacted. That is the virtuous (or vicious!) self-sustaining nature of the habit.

We can now go back to the original, albeit obscure and self-referential, intuition of roboticist Tim Smithers and his idea of the autonomy of behavior based on the "processes that, through interaction, are continuously forming the laws of interaction that can sustain and maintain the interaction needed to form them." These laws or norms are nothing other than the very conditions under which the habit is viable, that it can persist and sustain itself. This is a strong analogy with (metabolic) life that opens up the very possibility of *having a problem of your own* and *having to solve it*. The problem for the precarious, self-maintaining autonomous cell is persistence, avoiding decay and disintegration (see Barandiaran and Egbert, 2013 for a more detailed analysis). The same goes for the precarious, self-maintaining nature of a habit, and ultimately of autonomous sensorimotor life (see Barandiaran, 2008; Di Paolo et al., 2017). This approach also opens up a new mode of explanation that is characteristic of biological thinking and can be applied to sensorimotor dynamics: focusing on the nature and structure of constraints rather than the problems they are suppose to be adapted to solve. We return to these themes in the discussion section.

In what follows we introduce habit-based enactive robotics to illustrate and further develop the points we have briefly outlined. Inspired by SMC and enactive principles we build robots that are capable of generating spontaneously a complete ecology of habits that display structural constraints within the sensorimotor space. The results will help us discuss how enactive robotics can contribute to a new understanding of mind

and cognition with a deeper biological inspiration than what problem-solving can provide.

# 3. Model

## 3.1. Overview

The computational model simulates a two-wheeled robot that moves around a two-dimensional environment. The robot has two independently controlled motors; one for each of its wheels, allowing it to move forwards or backwards in a straight line or to turn in a variety of arcs, or on the spot. The robot's motors are controlled by an iterant deformable sensorimotor medium (IDSM) (Egbert and Barandiaran, 2014)—a habit-forming controller that is described in detail below.

The robot's environment is periodic in the sense that when the robot moves off one side of the environment, it appears on the opposite side. A stimulus source is also located in the environment. This source, which moves around the environment, can be thought of as simultaneously emitting a sound tone and a source of light, but in any given trial the robot is only capable of perceiving one of those sensory modalities (light or sound). We now present each element of the simulation in detail.

The robot and its environment are simulated using Euler forward integration with a time step of $\Delta t = 0.01$. Thus, a single time-unit consists of 100 iterations, and the IDSM is updated every iteration.

## 3.2. Stimulus

The stimulus source moves around the environment in a circle. Each rotation, it slows to a stop at its left-most position, before accelerating again to complete another rotation. This trajectory is specified by the following equations which describe the stimulus's position $(s_x, s_y)$ as a function of time $(t)$:

$$s_x = \frac{1}{2} + \cos\left(\frac{t}{10} + \sin\left(\frac{t}{10}\right)\right) \tag{1}$$

$$s_y = \frac{1}{2} + \sin\left(\frac{t}{10} + \sin\left(\frac{t}{10}\right)\right). \tag{2}$$

## 3.3. Robot

The simulated robot (Figure 1) has two independently controlled motorized wheels. It's position changes according to the following differential equations,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = k_s \cos(\alpha)(m_l + m_r) \tag{3}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = k_s \sin(\alpha)(m_l + m_r) \tag{4}$$

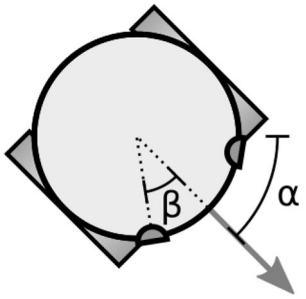



FIGURE 1
The simulated two-wheeled robot. The variable $\alpha$ specifies the orientation of the robot (direction of forward travel). The robot's sensors are located on its periphery with the parameter $\beta$ specifying the offset of the sensors from the direction the robot is facing.

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = k_s \frac{(m_r - m_l)}{2R}, \tag{5}$$

where $x$ and $y$ are the robot's position in the environment; $\alpha$ is its heading; $k_s = 0.25$ scales the speed of the motors; $R = 0.05$ is the robot's radius; and the variables $m_l$ and $m_r$ represent the velocities of the robot's left and right wheel motors.

We consider two robot sensor configurations. "Visual" robots have two directional sensors. The excitation of each, $V$, is the product of an attenuation factor due to distance from that sensor to the stimulus, and an attenuation factor due to misalignment between the orientation of the sensor and the relative direction of the stimulus. This second attenuation factor is calculated by taking the scalar product of a unit vector that points from the sensor to the stimulus and $\vec{o}$, a unit vector that specifies the direction that the sensor is facing.

Formally,

$$V = \underbrace{\frac{0.25}{0.25 + \|\vec{r}\|^2}}_{\text{distance}} \underbrace{\left(\frac{\vec{r}}{\|\vec{r}\|} \cdot \vec{o}\right)^+}_{\text{orientation}}, \tag{6}$$

where $\vec{r}$ is a vector that describes the position of the stimulus relative to the sensor; $\|r\|$ is the magnitude of that vector; and the $+$ superscript indicates that negative values within the parentheses are truncated to zero. The excitation of these sensors is thus highest when the sensor is close to the stimulus and directly facing it.

"Auditory" robots have two sensors that respond to the rate at which the sensor is approaching or moving away from the stimulus source. This is analogous to the Doppler effect whereby the perceived frequency of a sound when approaching is higher than when moving away from the listener. The excitation of an auditory sensor, $A$, is given by:

$$A = \frac{1}{2} + k \frac{\mathrm{d}||r||}{\mathrm{d}t} \tag{7}$$

where the first term can be thought of as the tones natural pitch which is offset by the relative speed of the sensor and the stimulus scaled by $k = \frac{3}{4}$ to keep the magnitude of the sensor (given the relative speeds of the robot and the stimulus) within a similar range of excitation values as simulated for the visual sensor.

To address the periodic boundaries of the environment, auditory sensors always use the nearest stimulus source as defined by the minimum image convention and visual sensors calculate the combined effect of 5 stimuli: one in the simulated space and four virtual copies of the sensor offset north, south, west and east of the simulated space arena by one arena width. This means that if, for example, a visual robot is close to the north boundary of the arena and facing north it can still see stimuli.

Visual sensors are offset from the direction the robot is facing by $\beta_v = \pi/5$ (see Figure 1). Auditory sensors are offset by $\beta_a = \pi/2$. The orientation of visual sensors is $\alpha \pm \beta_v$, i.e., perpendicular to the tangent of the robot's circular body at that position, facing outwards. Auditory sensors have no orientation.

## 3.4. IDSM

### 3.4.1. Overview

The IDSM is a robot controller intended to capture the idea of a self-maintaining pattern of sensorimotor activity. Inspired by the habitual behavior of people and by the enactivist concept of *"autonomous"* self-sustaining sensorimotor systems (see e.g., Di Paolo et al., 2017), the IDSM was designed such that patterns of sensorimotor activity reinforce the mechanisms that produce them (Egbert and Barandiaran, 2014; Egbert and Cañamero, 2014; Egbert, 2018). The IDSM has been used to explore how a habit-based individual can be trained to perform different tasks (Egbert and Barandiaran, 2014); how different forms of motor babbling can bias the subsequent formation of habits (Zarco and Egbert, 2019); how the essential variables of a biological autonomous system can be shared with the essential variables of a sensorimotor autonomous system (Egbert and Cañamero, 2014) and the extent to which IDSM-based sensorimotor autonomous systems can be considered to be adaptive (Egbert, 2018). In the present paper, we use the IDSM in a new way: to show, in a formal model, how sensorimotor-contingencies can play an essential role of sculpting the form of habits without themselves being explicitly internalized or represented by the "brain" or "controller" of an embodied agent.

The IDSM works by recording trajectories taken through sensorimotor space, i.e., how the sensorimotor state changes for various experienced sensorimotor states. When a sensorimotor state is experienced that is similar to one that has been experienced in the past, the motors of the robot are actuated in a way similar to how they were actuated in that previous experience. Memories of previous trajectories are gradually forgotten, unless they are reinforced by re-enactment and so the only patterns of behavior that can persist for long periods of time are those that are re-enacted. When the sensorimotor state is in an unfamiliar (or forgotten) state, motor activity is random. The IDSM used in this paper is very similar to that described in Egbert (2018). Any differences between the model here and that in Egbert (2018) are explicitly highlighted below.

A useful metaphor for understanding how the IDSM works is the paths that form on university campuses, where paths taken by students crossing a grassy field between academic buildings trample and kill the grass. The emergent dirt paths influence the trajectories taken by subsequent students, but the grass also regrows, so only emergent paths that are regularly traveled can persist in the long-term. This is essentially how the IDSM operates, but the trajectories taken and the paths that form are in sensorimotor space, rather than on a university campus. The dynamic also relates to the self-reinforcing nature of habitual behavior, where repeated performance of patterns of behavior (e.g., the direction you look when crossing the street; smoking a cigarette; or a tendency to worry) increases the likelihood of similar behavior being performed in the future. And to reiterate: the IDSM was designed to capture the enactivist concept of autonomy (a precarious self-maintaining system) in a sensorimotor system—see Di Paolo et al. (2017).

More formally, the IDSM can be thought of as a function, $f$, that transforms the robot's current sensorimotor state into an "output," i.e., the next moment's motor state: $f_t(S_t, M_t) \rightarrow M_{t+1}$. As this function is applied, the function itself also changes as a function of the current state of sensors and motors and the current state of the function: $\frac{\Delta \mathrm{f}}{\Delta \mathrm{t}} = g(f, S, M, \frac{\Delta S}{\Delta \mathrm{t}}, \frac{\Delta M}{\Delta \mathrm{t}})$. This change, which we shall now describe, was engineered so that sensorimotor state trajectories would bias the system to increase the likelihood that similar sensorimotor trajectories will be repeated in the future.

### 3.4.2. Tracking sensorimotor trajectories

As the robot's sensorimotor state changes, the IDSM maintains a set of records called *"nodes."* Each node describes the sensorimotor-velocity (i.e., the rate of change in all sensors and motors) for a particular sensorimotor-state at the moment that the node was created. Each node, $N$, is a tuple, $N = \langle \boldsymbol{p}, \boldsymbol{v}, w \rangle$, where $\boldsymbol{p}$ represents the sensorimotor state associated with the node (referred to as the node's 'position in sensorimotor space'); $\boldsymbol{v}$ indicates the sensorimotor velocity when the node was created; and $w$ indicates the weight of the node, a value that changes dynamically and is used to scale the overall influence of the node. We shall refer to these components using a subscript notation, where the position, SM-velocity vector, and weight of node $N$ are written as $N_{\boldsymbol{p}}$ and $N_{\boldsymbol{v}}$ and $N_w$, respectively.

As a robot controlled by the IDSM moves through sensorimotor states, new nodes are created recording the sensorimotor velocities experienced at different sensorimotor states. Specifically, when a new node is created, its $N_{\boldsymbol{p}}$ is set

to the current sensorimotor state; its $N_{\boldsymbol{v}}$ is set to the current rate of change in each sensorimotor dimension, and its initial weight, $N_w^0$.

The two vector terms ($N_{\boldsymbol{p}}$ and $N_{\boldsymbol{v}}$) can be thought of as existing within an abstract sensorimotor space (ASMS). A many to one mapping transforms any given position in the ASMS to a specific sensorimotor state, according to:

$$\begin{bmatrix} \Lambda(\mu_l) \\ \Lambda(\mu_r) \\ \sigma_l \\ \sigma_r \end{bmatrix} = \begin{bmatrix} m_l \\ m_r \\ s_l \\ s_r \end{bmatrix} \tag{8}$$

where $\mu_l, \mu_r, \sigma_l$, and $\sigma_r$ indicate a position in the four dimensional ASMS; the vector on the right indicates the sensorimotor state (i.e., the actual state of left-motor, right-motor, left-sensor, and right-sensor) associated with that ASMS position; and $\Lambda(x) = \sin[(4+x)2\pi x^3]$ is the non-linear function plotted in Figure 2. The purpose of this mapping is to avoid prescribing a characteristic rate of motor change, and instead to allow the IDSM to autonomously find habits with rates of motor change that are neither too fast nor too slow. Different regions of ASMS correspond to different rates of motor change. For instance, when abstract motor state $x \lessapprox 0.5$, a small change in that state variable corresponds to a small change in $\Lambda(x)$, the actual motor state (see the difference between the red circle and red X in Figure 2), where an equivalent change in $x$ when $x \gtrapprox 0.7$ corresponds to a greater change in the actual motor state (blue circle and X in Figure 2). When the sensorimotor state of the robot is unfamiliar, changes in motor activity are driven randomly (as explained below) and this $\Lambda$ mapping allows these random changes in sensorimotor space to correspond to slow or fast changes in motor state. By exploring different parts of the abstract sensorimotor state, the IDSM can experiment with sensorimotor patterns with different rates of change until ones that are self-reinforcing emerge. The ASMS is also treated as periodic so as to avoid the IDSM getting stuck at motor boundaries (as discussed in Egbert, 2018). ASMS state variables, $\mu_l, \mu_r, \sigma_l, \sigma_r$ all $\in$ $[0, 1]$ and the ASMS distance functions (described below) adhere to the minimum-image convention.

New nodes are added when the density of nodes near the current sensorimotor state is less than a threshold value, i.e., when $\psi(\boldsymbol{x}) < k_t$. Loosely speaking, $\psi$ is a measure of how 'familiar' the current sensorimotor state is, as estimated by summing a non-linear function of the distance (Equation 11) from every node with a positive weight to the current sensorimotor state. Formally,

$$\psi(\boldsymbol{x}) = \sum_N \omega(N_w) d(N_{\boldsymbol{p}}, \boldsymbol{x}) \tag{9}$$

$$\omega(N_w) = \begin{cases} 1 & \text{if } N_w > 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

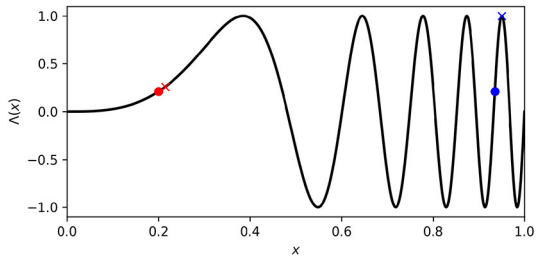
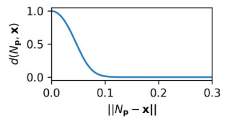



FIGURE 2
Abstract motor-state to motor-state mapping. Changes made to low abstract motor state [e.g., red circle and (X)] produce less change in actual motor state ($\Lambda(x)$) than when the abstract motor state is high (e.g., blue circle and X).

where $\boldsymbol{x}$ represents the current ASMS position, $k_t$ is a threshold parameter describing maximum node-density at which new nodes will be created and $d()$ is the following non-linear ASMS-distance function.

$$d(N_{\boldsymbol{p}}, \boldsymbol{x}) = \frac{2}{1 + \exp(k_d ||N_{\boldsymbol{p}} - \boldsymbol{x}||^2)} \tag{11}$$




### 3.4.3. Nodes influence the sensorimotor state

One time unit after a node has been created, it is added to the set of 'activated' nodes that influence the sensorimotor state according to:

$$\boldsymbol{m} = \frac{1}{\phi(\boldsymbol{x})} \sum_N \omega(N_w) d(N_{\boldsymbol{p}}, \boldsymbol{x}) (\underbrace{N_{\boldsymbol{v}}}_{\text{vel.}} + \underbrace{A(N_{\boldsymbol{p}} - \boldsymbol{x}, N_{\boldsymbol{v}})}_{\text{attraction}})^{\boldsymbol{\mu}} \tag{12}$$

This equation describes a weighted average of the influence of all of the nodes. The influence of each individual node is the sum of its "velocity" factor and its "attraction" factor. The velocity factor is simply the $N_{\boldsymbol{v}}$ vector (i.e., the sensorimotor velocity recorded when the node was created). The attraction factor is defined by

$$A(\boldsymbol{a}, N_{\boldsymbol{v}}) = \boldsymbol{a} - \left( \boldsymbol{a} \cdot \frac{N_{\boldsymbol{v}}}{||N_{\boldsymbol{v}}||} \right) \frac{N_{\boldsymbol{v}}}{||N_{\boldsymbol{v}}||} \tag{13}$$

and it causes the sensorimotor state to move toward the node. The attraction term is included to cause the system to move toward more familiar regions of sensorimotor space and to help stabilize patterns of repeated behavior (see Egbert and Barandiaran, 2014, 2015). The $\mu$ superscript in Equation (12) expresses that the IDSM only (directly) controls the motor components of the sensorimotor state. The sensory components are the result of the robot's relation to its environment and so are not directly controlled by the IDSM, but are, of course, *influenced* by the motor dynamics, indirectly through

the sensorimotor contingencies determined by the robot's environment and body.

The influence of each node is attenuated by a non-linear function of the distance between the node and the current sensorimotor state. This attenuation is expressed by the term $d(N_{\boldsymbol{p}}, \boldsymbol{x})$ and it means that nearby nodes affect the sensorimotor state and farther away nodes have little influence. The influence is also attenuated by a threshold function of weight, $[\omega(N_w)]$ such that only positively weighted nodes affect the motor state. Previous versions of the IDSM had a more complicated sigmoidal function in place of the simpler threshold function used here (Equation 10). Note that the degradation of the nodes and the threshold function of Equation (10) mean that when nodes are not visited for a long period of time they cease to have any influence whatsoever. Nodes that have degraded to this point essentially cease to exist.

After a node is created, its weight changes according to:

$$\frac{\mathrm{d}N_w}{\mathrm{d}t} = -k_{\downarrow} + k_{\uparrow} d(N_{\boldsymbol{p}}, \boldsymbol{x}) \tag{14}$$

In this equation, the first term represents a steady degradation of the node's influence and the second term represents a strengthening of the node that occurs when the current sensorimotor state is close to the node's position.

The influence of all nodes is summed and then scaled by the local density of nodes,

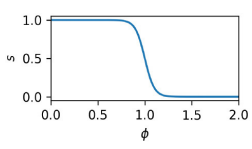
$$\phi(\boldsymbol{x}) = \sum_{N} \omega(N_w) d(N_{\boldsymbol{p}}, \boldsymbol{x}) \Big|_{\text{activated nodes}} \tag{15}$$

This equation looks similar to that used to calculate $\psi$ (Equation 9), but is different in that $\psi$ describes the local density of all nodes, where as $\phi$ describes the local density of activated nodes *only*.

### 3.4.4. Random motor activity

When the local density of activated nodes is low, motor behavior is random. This is accomplished by defining of a "switch" value, $s$, which specifies when the behavior is to be random and when it is to be controlled by the influence of the IDSM's nodes. The following equation expresses that $s$ is 1 when $\phi$ is low and 0 when $\phi$ is high; and that it moves between these values in a smooth, sigmoidal manner:

$$s = \frac{1}{1 + \exp(R_g(\phi(\boldsymbol{x}) - R_t))} \tag{16}$$




Here, $R_g$ and $R_t$ are parameters that specify the threshold of familiarity and the discreteness of the transition between random and non-random motor activity. This value is then used to switch between random motor activity ($\boldsymbol{r}$), and IDSM's output ($\boldsymbol{m}$), thus:

$$\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}t} = (1 - s)\boldsymbol{m} + s\boldsymbol{r}. \tag{17}$$

where the random motor activity vector, $\boldsymbol{r}$ is varied over time to produce a random walk in motor space as follows: every iteration, there is a $R_p$ chance that the components of $\boldsymbol{r}$ will be assigned random values selected from a normal distribution with a standard deviation of $R_\sigma$. The value of the $R_\sigma$ and other parameters can be found in Table 1.

## 4. Experiments and results

We now present two computational simulations of this model where we vary the sensory modality of the robot to explore how sensorimotor contingencies constrain the forms of the sensorimotor habits that can emerge and self-stabilize. The simulated robots and their environments are identical except that one robot's sensors are visual (as described above) and the other's are auditory. Formally, the only difference between these simulations is the equations that describe the stimulation of the robot's left and right sensors (Equations 6, 7).

To generate the data presented below, we ran 10 trials of each condition (i.e., we simulated 10 auditory robots and 10 visual robots) and for each condition, we selected a trial that displayed a wide variety of habits, and for which the simulated agent returned to one or more habits that it performed earlier but had stopped performing for some period of time. Not all trials did this—some instead rapidly fell into a pattern of behavior that was stable for the duration of the simulation. Data for all of the simulations is available at [DATA STORE LOCATION]. The analysis below covers visual simulation #9 and auditory simulation #0.

We now present an overview of the behaviors demonstrated by these two robots.

## 4.1. Visual sensors

The path taken by the visual robot through its environment is displayed in Figure 3A. This is a complicated and difficult to visualize trajectory, but it in fact involves several distinct repeated patterns of sensorimotor behavior (Figures 3B–H). To identify these, we first plotted the proximity of each of the IDSM's nodes to the robot's sensorimotor state over the course of the simulation (Figure 4). In this time series, it is easy to observe segments of time in which a particular set of nodes is repeatedly visited. For instance, when $750 \lessapprox t \lessapprox 1,200$, there are a few

TABLE 1 Parameters.

| Parameter | Value | Description |
|---|---|---|
| ROBOT | | |
| $R_s$ | 0.25 | Robot speed |
| $R_d$ | 0.1 | Robot diameter |
| $R_o$ | $\pi/5$ | Offset of sensors from center line of robot |
| $R_f$ | 0.25 | Falloff of sensor excitation with distance |
| LIGHT | | |
| $L_v$ | 0.1 | Light velocity scale constant |
| $L_r$ | 1/3 | The radius of the light's circling motion |
| IDSM | | |
| $k_d$ | 500 | Fall-off of non-linear ASMS distance measure |
| $k_\downarrow$ | 5 | Node weight decay rate |
| $k_\uparrow$ | 2000 | Node-reinforcement rate constant |
| $k_t$ | 1 | Node-density threshold parameter, influencing when new nodes are to be added |
| $N_w^0$ | 1000 | Node weight upon creation |
| MOTOR BABBLING | | |
| $R_p$ | 0.1 | Chance of random reassignment of random-motor walk direction and velocity |
| $R_\sigma$ | 0.25 | Standard deviation of random motor components |
| $R_t$ | 1 | Node-density threshold parameter, influencing when motor activity is random |
| $R_g$ | 20 | Steepness of transition between random and IDSM driven motor activity |

nodes with indices close to 5,000 that are repeatedly visited (tan horizontal sequence of dots), and the same set of nodes are briefly revisited three times in the final 500 time units of the simulation. We manually identified these repeated patterns of sensorimotor activity, and assigned each pattern a label and a color (the example just provided is labeled "D" and colored tan). Times when the robots behavior is not clearly repeated were not assigned a label and are marked with a light gray color. To be clear, repetitions were identified in sensorimotor space (*via* the identification of repeatedly visited nodes), not in physical space. Returning to Figure 3, we can see that when the robot is performing the D pattern, it is moving around the environment, regularly turning in loops with squarish corners. Figure 5 shows how the state of the sensors, motors and the distance between the stimulus and the robot as the trial progresses. We can see that each of the colored regions tend to occupy particular regions of the sensorimotor space (the sensorimotor 'habitat' of the habit—see Buhrmann et al., 2013) and certain patterns of behavior involve the robot pursuing the moving light (e.g., F-behavior) while others avoid it (e.g., D-behavior).

Each of patterns B–H can be construed as a potential habit, i.e., a pattern of autonomous, self-maintaining sensorimotor activity. To justify this claim, we observe that

1. A node can persist for an extended time if and only if it is regularly "visited," i.e., if and only if the robot's sensorimotor state regularly comes sufficiently close to the node such that the node's weight always kept above zero.
2. The robot's sensorimotor trajectory depends upon the nodes in that (a) the nodes directly control how the motor components of the sensorimotor state change, and (b) the nodes indirectly constrain and influence the sensory components *via* their effect upon the motors.

From these observations it follows that the nodes enable the very thing that they depend upon to persist and reflexively, the enactment of the pattern of behavior enables the very thing that it depends upon (the nodes) for its persistence. Similar to a model of biological autopoiesis (Varela et al., 1974), where inherently unstable components such as metabolism and membrane are mutually enabling, the components of these patterns of behavior in this model are in a relationship of mutual interdependence and support, and thus can be considered autonomous under some readings of the enactivist literature. The behavior itself is operationally closed entity (a "unity"), constituted by inherently unstable components (the nodes and the sensorimotor trajectory) and yet persists thanks to its own enactment or performance.

More detailed analysis of the operational closure of these systems is outside the scope of the present paper (but see Egbert, 2018 for initial analysis of the precarious autonomy of a simple IDSM-based habit). We can see the basic idea however, when we consider how the weight of the nodes changes as time passes in the simulation (see Figure 6). Recalling that when the weight of a node reaches zero, that node ceases to exist, it is clear that only nodes that are regularly visited can persist in the long term. For a collection of nodes to be visited, the sensorimotor trajectory must move in a particular trajectory and the sensorimotor trajectory is largely determined by the activity of nodes. It follows from this that the only way that a pattern of behavior can persist is if it is one that causes the repeated revisitation of its constituent nodes. We can also note that this self-reinforcement of a habit need not be constant or contiguous. For example, the B-habit (red) is established early in the simulation [$t \approx 250$)] and then is not visited again until $t \approx 1,250$, where it is enacted a few times and the nodes are reinforced such that they survive until close to the end of the simulation. The parameters that prescribe the rates of node weight reinforcement ($k_{\uparrow}$) and degradation ($k_{\downarrow}$) determine how regularly patterns of behavior must be enacted if they are to persist.

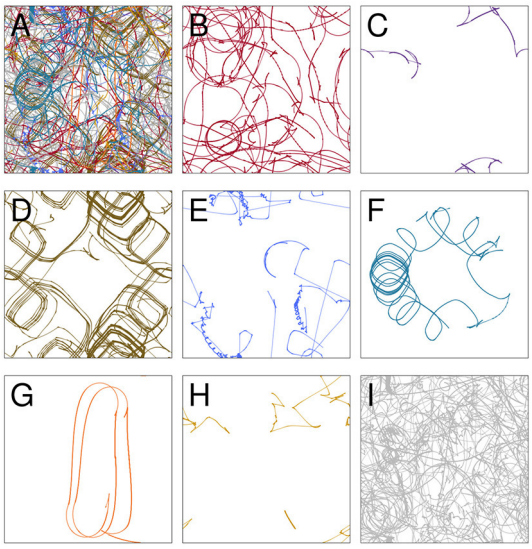



FIGURE 3
The trajectory taken by the visual robot **(A)**, broken down into 8 patterns of repeated sensorimotor activity, i.e., habits **(B–H)**. The portions of the trajectory where the robot's behavior is not clearly repetitious are included in plot **(I)**. Each of the nine plots shows the full 1 × 1 arena. Video showing this trial is available at https://www.youtube.com/watch?v=2v1TyvKz9qw.

## 4.2. Auditory sensors

We performed a similar analysis on the auditory robot. Figure 7A shows the complete path taken by the auditory robot in its environment, broken down by habit (Figures 7B–H). Figure 8 shows the weight of each node which also provides

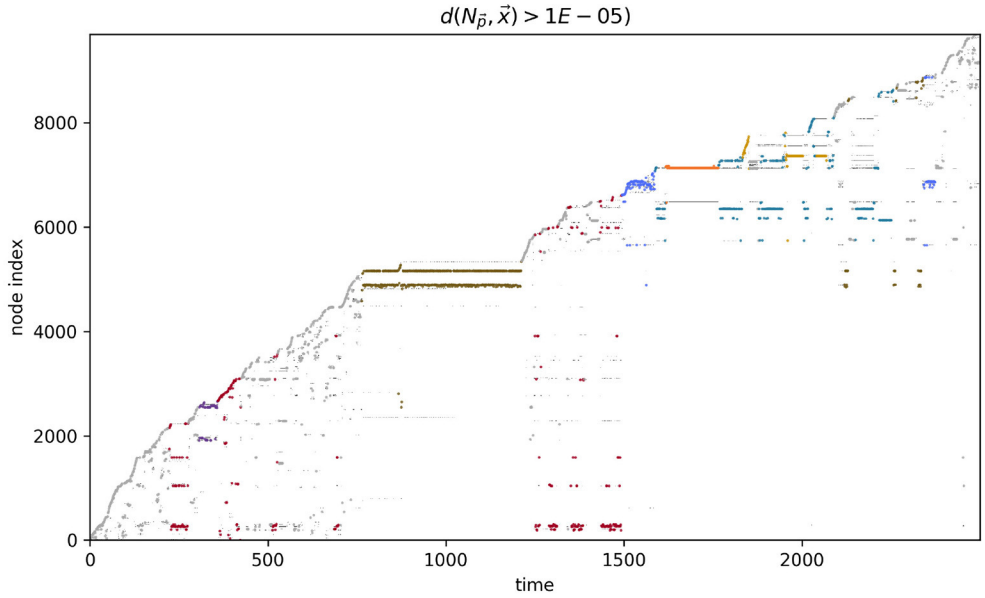



FIGURE 4
Node proximity during the course of the visual robot's simulation. The darker the point in this plot, the closer that node is to the robot's sensorimotor state at that point in time. The larger colored points indicate the index of the node that was closest (in sensorimotor state) to the robot's sensorimotor state at that time. The color of the closest-node points indicates the habit that those nodes are associated with. These colors match those in Figure 3.

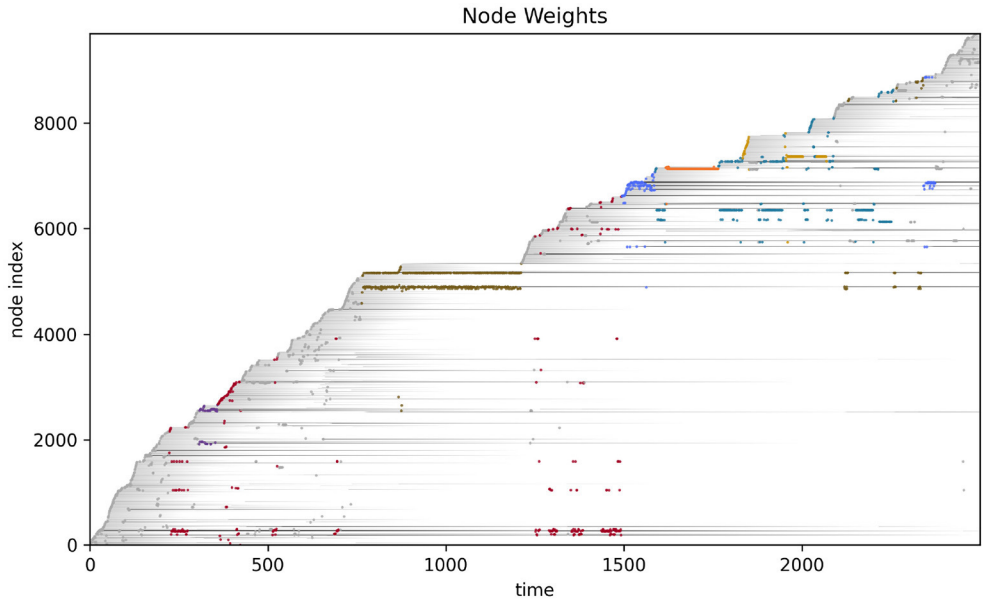



FIGURE 5
State of sensors and motors for the simulation of the visual robot.

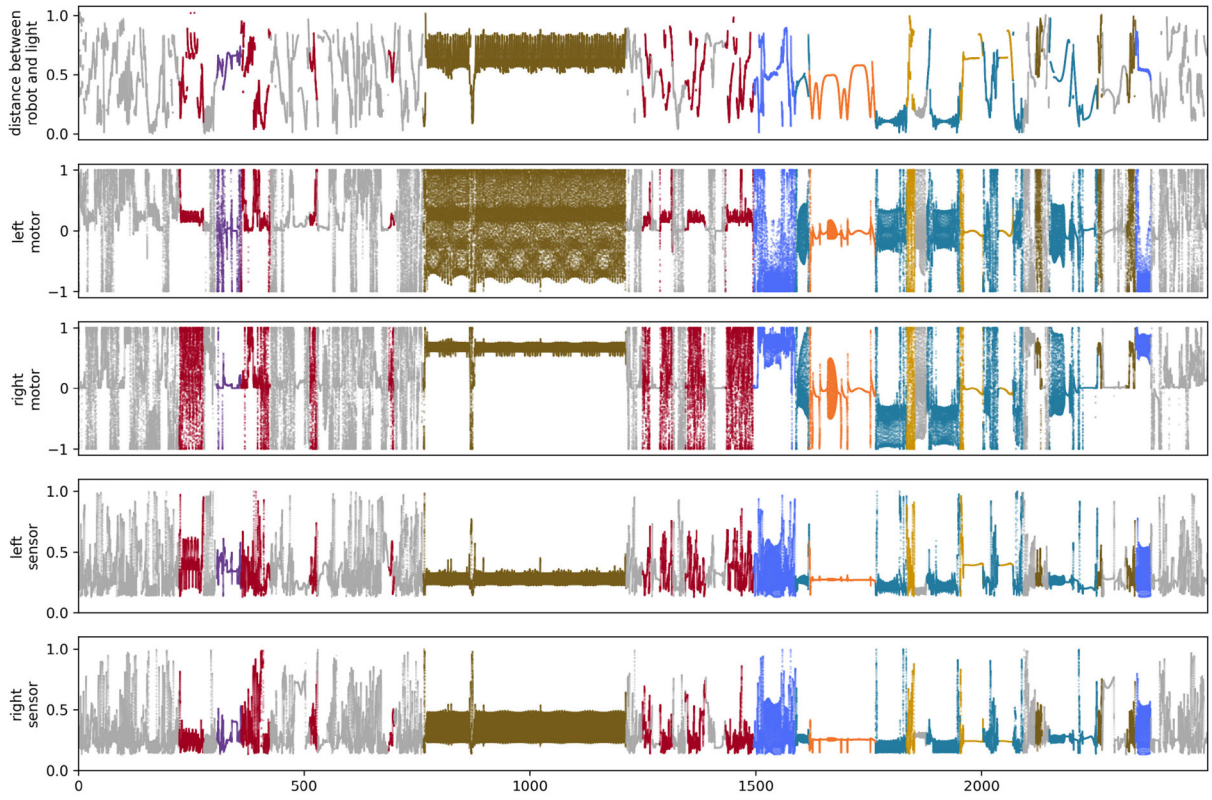



FIGURE 6
Node weight during the course of the visual robot's simulation. The weight of each node (shades of gray) are shown as time progresses during the simulation. After most nodes appear, their weights decay (become lighter as time passes), but some nodes are regularly visited and their weights are reinforced (long gray horizontal lines). The closest node at any time is plotted in color, just as in Figure 4.

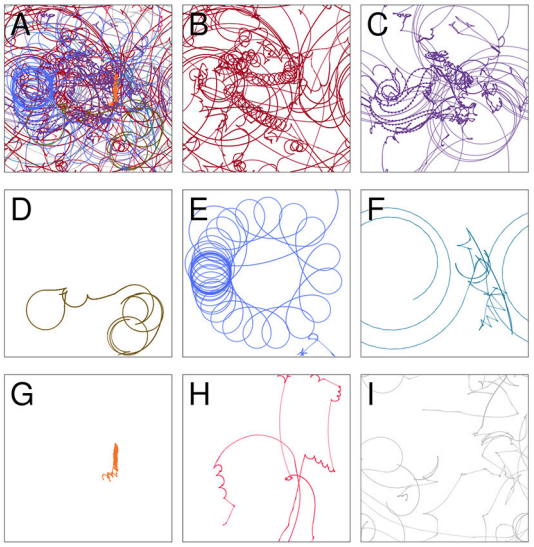



FIGURE 7
Spatial trajectory taken by the auditory robot **(A)**; broken down into "habits" i.e., self-reinforcing patterns of behavior **(B–H)**; and portions of the trajectory not associated with any habits **(I)**. Each of the nine plots shows the full 1 × 1 arena. Video showing this trial is available at https://www.youtube.com/watch?v=lf_WeclEtCM.

an indication for how close each node is to the current sensorimotor state of the robot and the timeseries plots for the auditory robot are shown in Figure 9. It is worth noting that though these patterns of behavior may seem random, they all (both visual and auditory) relate to the moving stimulus source in regular, non-random ways. The regularities in the interaction with the stimulus are much more easily seen in the animations linked to in the captions of Figures 3, 7).

## 4.3. Comparison of sensorimotor structures

Even when two behavior's functions are the same (e.g., they both accomplish taxis) the forms of the underlying sensorimotor habit can be different. As a case in point, we can compare auditory habit E with visual habit F. In both cases, the robot circles around the stimulus, maintaining a similar approximate distance from the stimulus as the stimulus moves in spurts around the environment. When we look at the two patterns of sensorimotor activity (Figure 10), we see two different pictures. Each row of this visualization shows the spatial trajectory of the robot (left) followed by four projections of the 4D sensorimotor trajectory of the robot and it is obvious that the two sensorimotor trajectories are qualitatively different. The auditory pattern involves less diversity in the range of states visited; the mean and other statistical properties of the sensorimotor trajectories are also clearly different, etc. But every habit in this model is unique—how much of the difference between these habits is simply due to the fact that they are different habits, and how much is due to the difference in the robot's sensorimotor modalities?

To address this question, we can look at projections of the sensorimotor trajectories of all of the habits of the visual robot (Figure 13) and observe that there are regularities in the forms of the habits of this robot. This is most readily seen in the left-sensor vs. right-sensor projection (upper right) which reveals elements of this robot's sensorimotor contingencies, including aspects of both its sensorimotor habitat and environment (Buhrmann et al., 2013) of the robot. Specifically, the bilateral symmetry of the robot's embodiment result in a (statistical) mirror symmetry across the left sensor = right sensor diagonal. The directionality of the visual sensors and their different orientations mean that it is impossible (in the environment with a single stimulus) to maximally stimulate both left and right sensors concurrently, leading to the L-shape apparent in upper-right plot of Figure 13, with arcs that pass between the stems of the L as the robot turns such that the stimulus moves from being in front of one sensor to in front of the other. These features are also found in the other nine visual robot trials.

Similar plots for the auditory robots (Figure 14) also reveal regularities. Like the visual robot, the auditory robot's embodiment is bilaterally symmetric and so it has the same symmetry across the diagonal in left-sensor/right-sensor plot, but the non-directionality of the auditory sensors and the fact that their excitation does not fall off with distance make other aspects of the sensorimotor activity different. The auditory trajectories tend to repeat patterns of diverging from a center point in short arcs in a pattern of returning and diverging again, in a different direction. These dynamics can be seen in greater detail in Appendix A, which shows the projection in sensorimotor space of each habit individually.

Sensors and motors are involved in a recursive relationship of influence, where sensors influence motors (*via* the controller) and motors influence sensors (*via* the effect of actions upon the environment and the agent's relationship to the environment). It therefore would make sense that the constraints imposed upon sensory dynamics by different sensory modalities would constrain the motor dynamics. In other words, under one sensory modality, certain motor trajectories will be more readily repeated (and thus stabilized) than others, and which motor trajectories are more readily repeated would depend upon the sensory modality.

We do, in fact, see differences in the distributions of motor states between the two sensory conditions (Figure 11). Visually, we can see that the distributions of motor (and abstract motor) states in the visual robots seem different than that of the auditory robots (Figure 11); and the distributions of motor states within these groups in independent runs of the visual trial seem more similar to each other (Figure 12) than they are

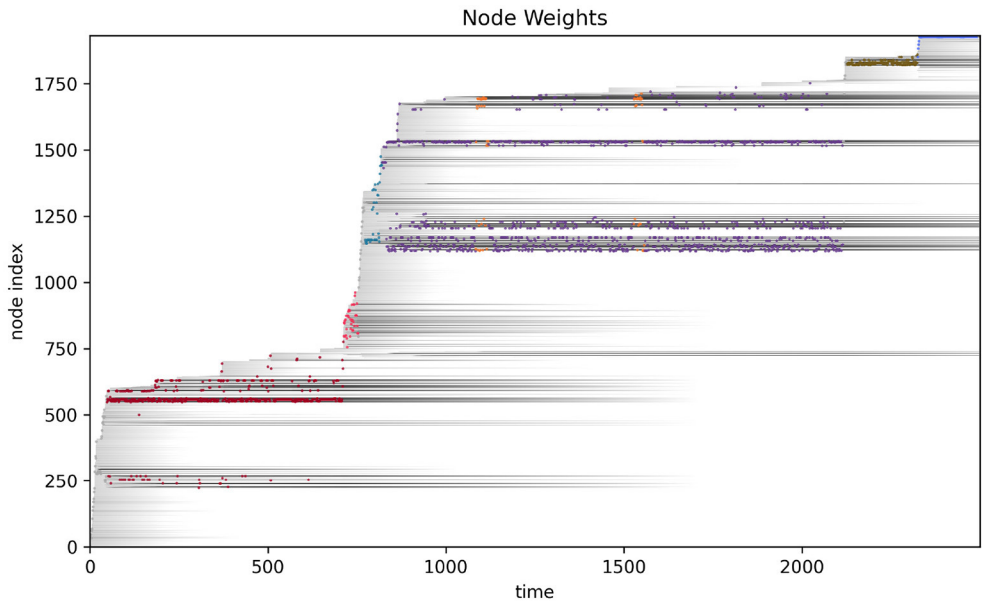



FIGURE 8
Node weights with nodes closest to the current sensorimotor state of the auditory robot highlighted and marked by color of habit.

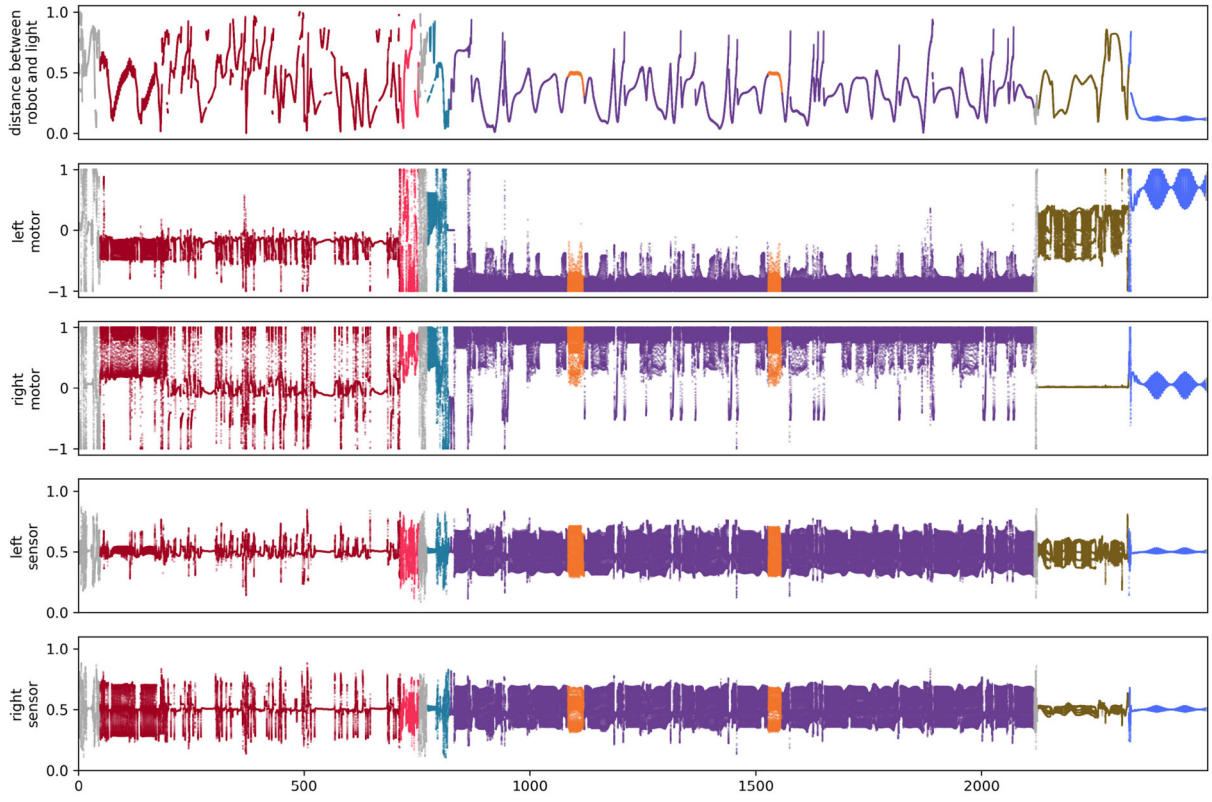



FIGURE 9
State of sensors and motors for the simulation of the auditory robot.

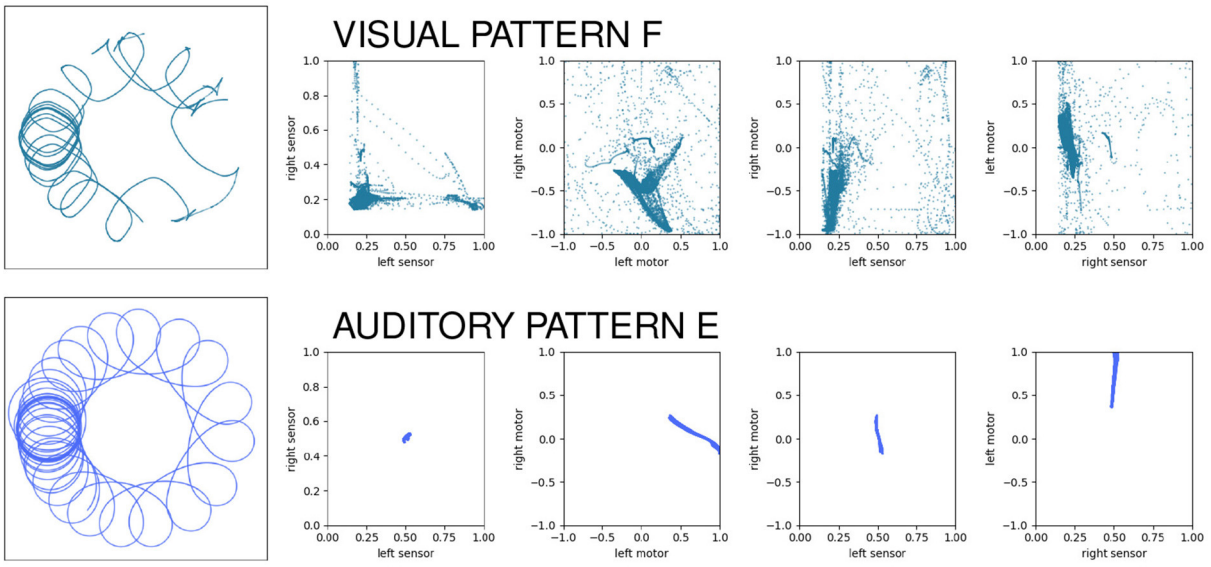



FIGURE 10
Contrasting the sensorimotor dynamics of two "circling" habits. In both cases, the robot circles the stimulus, but they use different sensory modalities to do so. The patterns of sensorimotor activity are clearly different despite the external "functional" description of the behavior.

to the other group. We confirmed the statistical significance of these differences by using Kullback–Leibler divergence to assess the distributional conformity of 50 visual trials and 50 auditory trials. The within-group (intra-modal) distributions were significantly more similar than the between-group (inter-modal) distributions for $m$ ($t = -4.67$, $p = 0.0114$) and for $\mu$ ($t = -8.507$, $p = 0.0215$).

In the absence of the self-reinforcing dynamic of the IDSM, habits, motor activity would be random, in which case we would expect the same distribution of motor states in both sensory modalities. The histograms in (new) (Figures 11, 12) show that this is not the case, and thus provide support for the claim that the sensorimotor contingencies implicit in the different sensory modalities constrain not only the sensory dynamics, but also motor dynamics—and in doing so, they play an important role of constraining the form of emergent autonomous sensorimotor dynamics.

The primary point that we wish to communicate using this model is that sensorimotor contingencies influence which patterns of autonomous sensorimotor behavior can emerge and persist. It is worth emphasizing that the plots in Figures 10, 13, 14 are not plots of the sensorimotor contingencies of the robot—they are plots of the robots' self-maintaining habits. Nevertheless, we see in these plots the *influence* of sensorimotor contingencies—the way that the contingencies have constrained the set of habits that can form and self-stabilize, producing the regular patterns described in the paragraphs above.

The two different sensory modalities entail two different sets of sensorimotor contingencies. The different sensorimotor contingencies each imply a different set of possible self-stabilizing habits, and so the habits that form with one sensory modality are more similar to each other than they are to the habits that form with the other sensory modality. Two *visual* habits will be different from each other, but they will generally be more similar to each other than auditory habits. This is what we mean by the claim that the form of sensorimotor habits are constrained by their sensory modality and it is apparent in the regularities described above and seen in Figures 13, 14. In this account, sensorimotor contingencies play an important role. They influence the form of the emergent autonomous sensorimotor strutures, but the rules themselves are not internalized or learned in any other way.

## 5. Discussion

### 5.1. Solving somebody else's problems vs. having your own problems

The robots we have just presented display a number of interesting and challenging features for traditional problem-solving centered artificial intelligence (biologically inspired or otherwise). We see in this model how sensorimotor dynamics can have problems of their own. As we have seen AI and robots more specifically are generally built to solve problems. But whose problems? Perhaps the engineers' problems or those of their clients, perhaps animal problems that require robotic models to be better understood, maybe societies problems but certainly

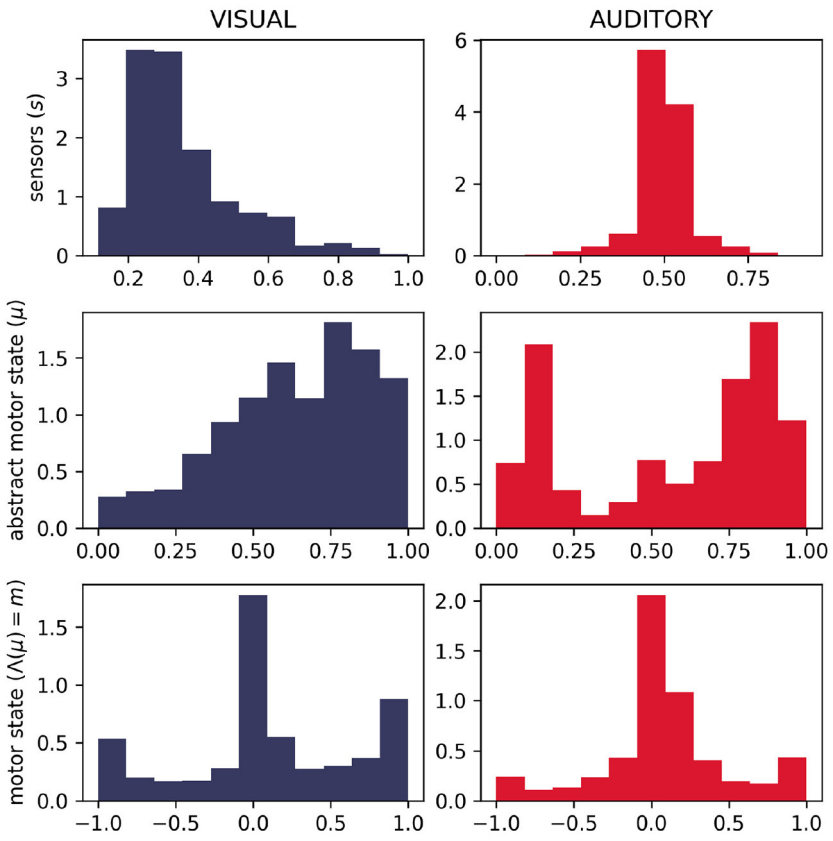



FIGURE 11
Histograms showing difference in distributions of sensorimotor states for 50 visual and 50 auditory trials. As expected, the different sensory modalities produce different distributions of the sensor state. It can also be seen that in visual trials, the abstract motor-state tended to be higher (corresponding to faster rates of motor state change—see Figure 2) and the motor state tended to be more extreme.

not those of the robot itself. It is thus important to distinguish between having one's problem vs. being able to solve a problem. A bacterium that swims up a gradient toward the resources it needs to survive is doing both. It has a problem (it needs food to persist) and it solves that problem by navigating through its environment. AI, on the other hand, is capable only of the former. A "self-driving" car that takes me to a restaurant dinner is solving a problem (my need for food), but that problem is mine, and not the car's.

The enactive robots presented in this paper where not designed to do anything, to solve any problem. None of our analysis was described in terms of success or failure at the performance of a task. The architecture makes habits emerge and with them a basic sense of having a problem: the habit *needs* to enact itself in order to persist. The norms of behavior here are thus only in terms of the persistence of the behavior itself (and concurrently the mechanism that produces it). Surely, life is full of problems, we certainly have enough with our own and don't need to be concerned with creating new ones . . . unless we want to understand what it is to have a problem and how life is, itself, a source of problems. Enactivism embraces a deep conception of life as the self maintenance of a network of precarious processes. In this sense life is inherently a source of problems. By making it possible for habits to emerge and sustain their own existence through recurrent sensorimotor interaction we have shown how this conception of life can be practically transferred to a "way of life" in sensorimotor robotics.

A research program on enactive robotics can and must address how sensorimotor life is a source of problems. We can move beyond the single habit and envision how the network of habits that constitutes the agents can display emergent problems of its own: that of keeping the whole network alive, coping with variations on the way the environment affords or precludes to enact them. This line of enquiry requires further work but the present model requires little development to start addressing it.

## 5.2. Constraints and form in biological and behavioral explanation

Biological inspiration in robotics can also move beyond the analogy between metabolic or physiological life and its

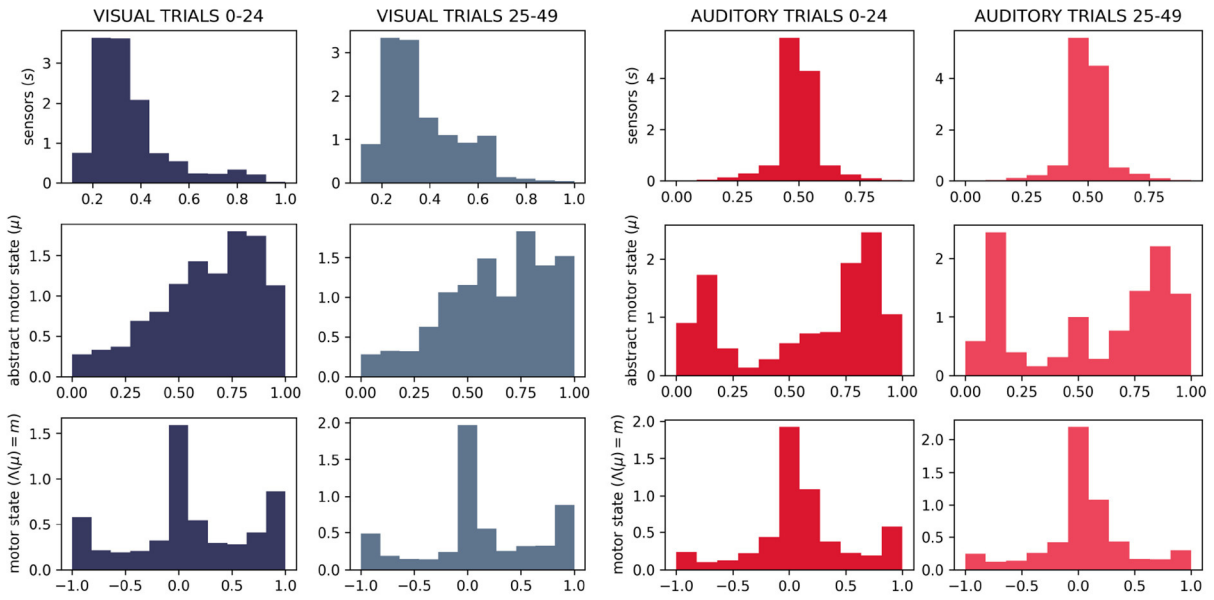



FIGURE 12
Histograms showing in-group similarity in sensorimotor state for visual trials **(left)** and auditory trials **(right)**.

sensorimotor counterpart. In particular the models presented and analyzed in this paper show how sensorimotor life and the problems that emerge within are constrained by sensory embodiment and how such constraints constitute an important part of what it means to explain life.

Whether it is understood as more fundamental-than, complementary-to or directly at odds-with (Newman, 2018) natural selection, it is the repeated, regular, robust and often phylogenetically independent appearance of forms or structures in biology what becomes in itself an object of study that cannot be reduced-to or deduced-from life as a response to environmental problems. In this sense biologically inspired enactive robotics can learn from biological development and its organization and import into psychology some explanatory strategies followed by biologists. Darcy Thompson, René Thom, Pere Albert, etc. all conceived biological forms to be both *explananda* and *explanans* on themselves. Homologies, analogies, inherency, are all concepts directed to capture the re-emergence of certain forms (and functions!) in biological organisms. The notion of constraint plays here a fundamental role. The way in which different layers of materiality and self-organization limit and channel the emergence of viable forms is essential to biological explanation.

If biological inspiration within the problem-solving paradigm concluded that the material embodiment of robots permits to offload and transform the problems to be solved, the biological inspiration of the enactive paradigm can conclude that the embodiment loads and informs the problems that constitute the sensorimotor agent. As we have seen, sensory modalities constraint motor trajectories, which in turn, shape how habits get stabilized. Thus, in a way that parallels the explanatory role of constrains in evolution, we can hypothesize that behavioral variability is not free (to explore potential solutions to cognitive problems) but is constrained and channeled by the embodiment of the agent.

Now, before the advent and widespread influence of evolutionary theory in biology, the latter synthesis with molecular biology and subsequent expansion to psychology, sociology and even epistemology, or philosophy more generally, the concept that was key for the continuity between life and mind was that of habit.

## 5.3. Enactive robotics revisited

There are different takes on how enactivism translates into robotics. As mentioned in the introduction it is possible to simply reject representational functionalism, or to claim for the importance of embodiment, or to demand that robot be endowed with living bodies, or to introduce some feedback mechanisms that parallel those provided by emotion bearing bodies. We have taken a different approach—that envisioned by Smithers (1997) and latter developed by Di Paolo (2003), Barandiaran and Moreno (2006), and Barandiaran (2008). What a research programme in enactive robotics entails is the study of the organization of sensorimotor life: the form of viability of different habits, the topology of the network of habits that

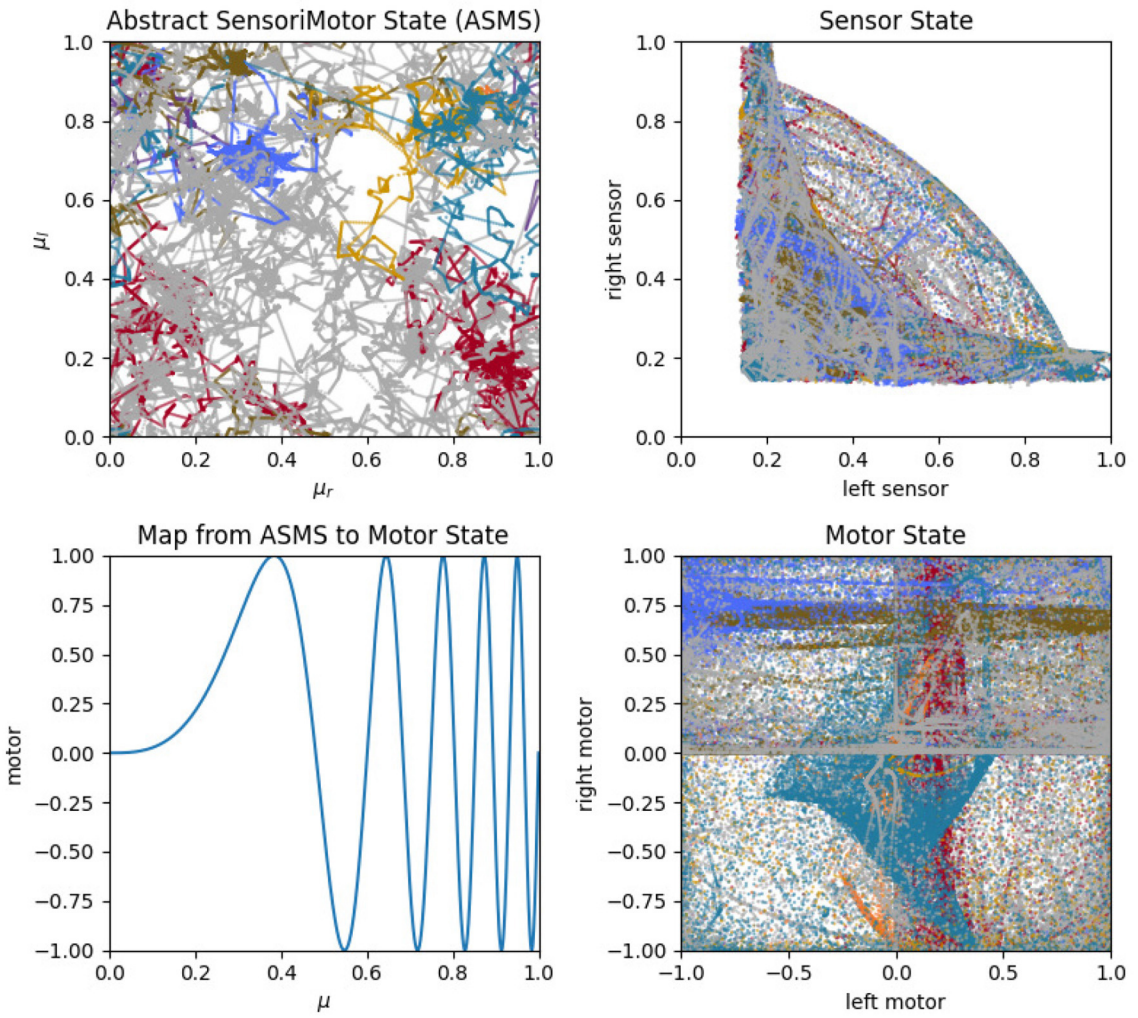



FIGURE 13
Sensorimotor dynamics for the visual robot. These plots show motor trajectories in abstract sensorimotor space **(upper left)**; two projections of the actual sensorimotor trajectories of the robot **(upper** and **lower right)**; and the mapping function, $\Lambda$, that transforms the abstract motor state to an actual motor state (see Section 3.4.2). Colors indicate habits as in previous plots.

unfolds over development, the shape of the habitat that is thus constructed, the structure of the world that is experienced. Habit-based robots as designed here are capable of individuating habits and of creating an ecology of habits that can easily be understood as a form of sensorimotor self.

Not only has enactivism informed robotics (Ziemke and Lowe, 2009; Vernon, 2010) but robotics has often served enactivism (Beer, 1995; Di Paolo, 2003; Aguilera et al., 2016) by clarifying its claims, pushing theoretical development, operationalizing its concepts or penetrating diverse problems. The model we developed here can be aligned with the latter. It can be used to clarify and make explicit the often obscure original formulation of enactivism that "cognition in its most encompassing sense consists in the enactment or bringing forth of a world by a viable history of structural coupling" (Varela et al., 1991, p. 205). We have shown how a robot endowed with an IDSM can bring into being a number of habits, the world that it brings forth is the habitat, more specifically, the structured set of sensorimotor contingencies that the agent inhabits or enacts. This habitat must be viable in the sense that habits must be sustainable and results from a (developmental) history of sensorimotor (structural) coupling. And in so doing, we have demonstrated how sensorimotor contingencies can *directly* constrain or "sculpt" the form taken by sensorimotor habits without requiring any virtualization, i.e., without any internal model or representation of the sensorimotor contingency.

Perhaps it is worth clarifying why the nodes do not constitute an internal model or representation. The nodes do not stand-for something else other than the sensorimotor dynamics they partake in. It is not possible to operate upon the nodes in a decoupled offline mode that is not itself the enactment of a behavior, and there is not additional module or subsystem

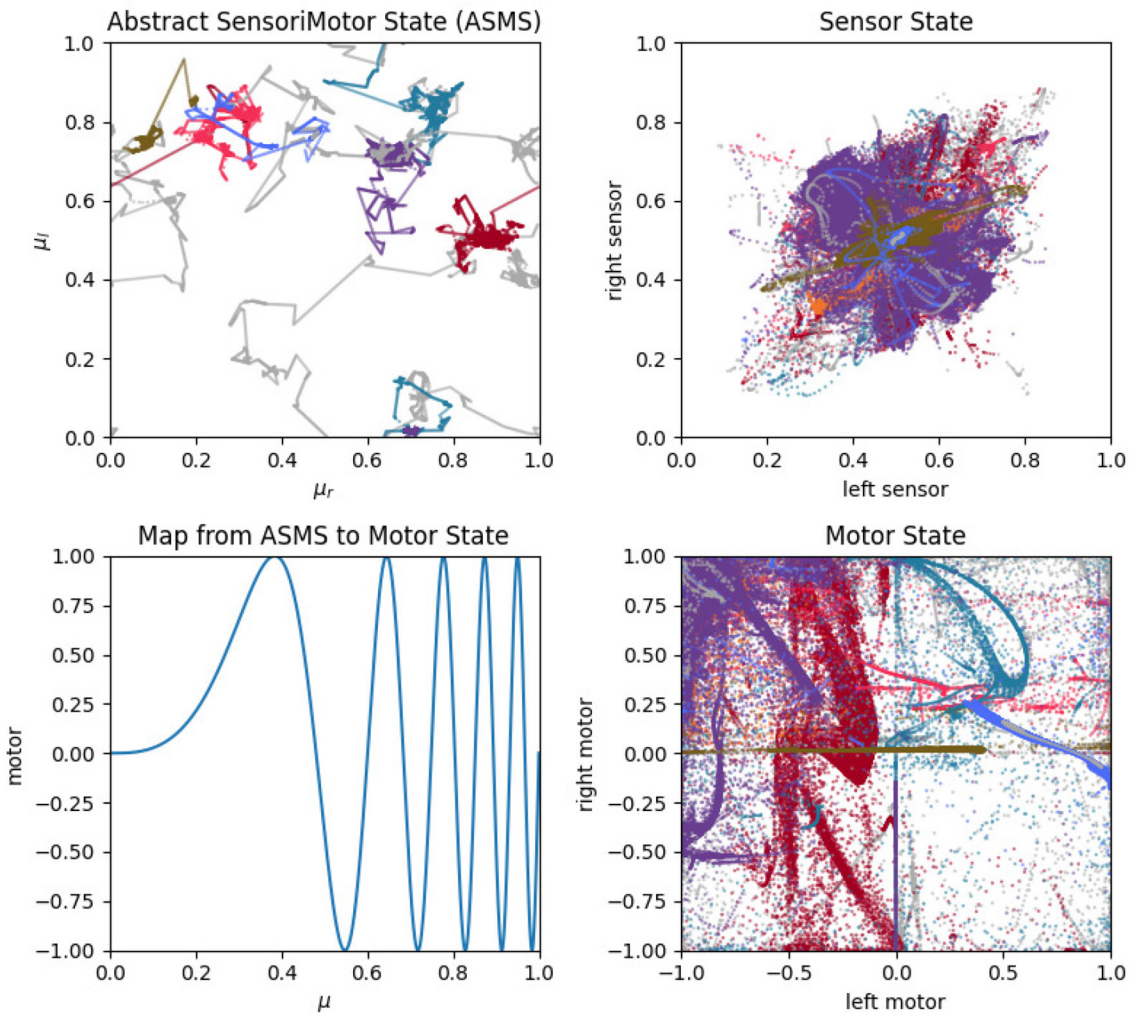



FIGURE 14
Sensorimotor dynamics for the auditory robot. This figure is the same as Figure 13 but plotted for the auditory robot.

"consuming" such nodes to carry out any further operation. In these ways, the nodes do not represent a behavior or habit, but they embody it—they constitute the habit *together* with elements of the robot's bodily and environmental dynamics.

## 6. Conclusion

A central focus across the cognitive sciences is upon problem-solving ability and tremendous progress has been made in understanding how to mechanize problem solving. Much of AI and robotics research is validated by how well some artifact (neural network, human being, robot, etc.) performs at a problem-solving task (chess, maze navigation, bipedal walking, etc.).

However, the conflation of "problem-solving ability" with all of the phenomena associated with 'being a mindful body leaves out a number of features that demand be put at the center of (enactive) theorizing: historicity, embodiment, habitat, precariousness, identity, norm-establishing, etc. All these dimensions of mental life are worthy of study and remain outside of the problem-solving frame that scaffolded the development of Artificial Intelligence and Robotics.

We have here presented a set of robots with different sensory modalities that spontaneously develops a complex ecology of sensorimotor habits. These are constrained by the sensory modality of the robot and give rise to sensorimotor habitats of specific forms. The very nature of habits thus developed, understood as self-sustaining forms of sensorimotor activity, has shown how robots must first have their own problems instead of solving those posited by external observers; and, that, in doing so, they must assert a form of life whose structure and topology must be taken as the object of study. In particular we have seen how the form of sensory embodiment shapes potential sensorimotor contingencies and these constraint the shape and type of sensorimotor habits that emerge during development.

To be alive is not a computable function but the way in which materiality (implementation), behavior and function are deeply intertwined. What was once claimed as the triumph of functionalism as the clear conceptual separation between matter, behavior and machine state transitions is now its deepest weakness. The enactive approach brings all three together again. To be fair it is not the materiality of the robot's "body" that is a stake here (neither does the simulation itself possess any body beyond the computer in which the simulation was carried out, not the physical body, e.g. wheels and sensors, would be at stake was the robot to be implemented in real life), but the materiality of the sensorimotor mapping, its precarious existence, its fading structural stability.

Perhaps the "artificial sciences" (AI, artificial life, robotics, exploratory computational modeling, etc.) would benefit from similarly investing more time in targets other than problem-solving ability. Biologists are sometimes accused of suffering "physics envy"—i.e., wishing that the objects of their study were more easily and completely summarized by simple, provable equations. Perhaps we enactivist and embodied researchers can be accused of "problem-solving envy" a desire for our artifacts and theories to be equally or more capable of solving problems as the expert systems or disembodied neural networks of other problem-solving focused approaches. And perhaps this envy is a distraction, and impeding our progress toward understanding minds. In fact, robotics cannot only reveal itself as an engineering practice directed at solving problems but also as a philosophical practice aimed at posing the right problems. This is a contribution that enactive robotics is ready to do.

## 7. Permission to reuse and copyright

Figures, tables, and images will be published under a Creative Commons CC-BY license and permission must be obtained for use of copyrighted material from other sources (including re-published/adapted/modified/partial figures and images from the internet). It is the responsibility of the authors to acquire the licenses, to follow any citation instructions requested by third-party rights holders, and cover any supplementary charges.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

ME conceived, implemented, ran simulations and experiments, and authored the manuscript. XB contextualized research, interpreted results, and authored the manuscript. Both authors contributed to the article and approved the submitted version.

## Funding

This work was supported in part *via* funding from the Digital Life Institute, University of Auckland. XB acknowledges funding from the Spanish Ministry of Science and Innovation for the research project Outonomy PID2019-104576GB-I00 and IAS-Research group funding IT1668-22 from Basque Government.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Aguilera, M., Bedia, M. G., and Barandiaran, X. E. (2016). Extended neural metastability in an embodied model of sensorimotor coupling. *Front. Syst. Neurosci.* 10, 76. doi: 10.3389/fnsys.2016.00076

Ashby, W. R. (1952). *Design for a BRAIN: The Origin of Adaptative Behaviour, 2nd Edn*. London: J. Wiley.

Barandiaran, X., and Moreno, A. (2006). On what makes certain dynamical systems cognitive: a minimally cognitive organization program. *Adapt. Behav.* 14, 171–185. doi: 10.1177/1059712306014 00208

Barandiaran, X. (2007). "Mental life: conceptual models and sythetic methodologies for a post-cognitivist psychology," in *The World, the Mind and the Body: Psychology after Cognitivism*, eds B. Wallace, A. Ross, T. Anderson, and J. Davies (Exeter: Imprint Academic), 49–90.

Barandiaran, X. E. (2008). *Mental Life. A Naturalized Approach to the Autonomy of Cognitive Agents* (Ph.D. thesis). University of the Basque Country, Spain.

Barandiaran, X. E. (2017). Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. *Topoi* 36, 409–430. doi: 10.1007/s11245-016-9365-4

Barandiaran, X. E., and Egbert, M. D. (2013). Norm-establishing and norm-following in autonomous agency. *Artif. Life* 20, 5–28. doi: 10.1162/ARTL_a_00094

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artif. Intell.* 72, 173–215. doi: 10.1016/0004-3702(94)00005-L

Buhrmann, T., Di Paolo, E. A., and Barandiaran, X. (2013). A dynamical systems account of sensorimotor contingencies. *Front. Psychol.* 4, 285. doi: 10.3389/fpsyg.2013.00285

Buss, D. M. (1998). *Evolutionary Psychology: The New Science of the Mind*. London: Allyn and Bacon.

Canaan, R., Salge, C., Togelius, J., and Nealen, A. (2019). "Leveling the playing field: fairness in AI versus human game benchmarks," in *Proceedings of the 14th International Conference on the Foundations of Digital Games, FDG '19* (New York, NY: Association for Computing Machinery), 1–8.

Cliff, D., Husbands, P., and Harvey, I. (1993). Explorations in evolutionary robotics. *Adapt. Behav.* 2, 73–110. doi: 10.1177/105971239300200104

Damiano, L., and Stano, P. (2021). A wetware embodied AI? Towards an autopoietic organizational approach grounded in synthetic biology. *Front. Bioeng. Biotechnol.* 9, 724023. doi: 10.3389/fbioe.2021.724023

Di Paolo, E. A. (2003). "Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop," in *Dynamical Systems Approach to Embodiment and Sociality*, eds K. Murase and T. Asakura (Adelaide, SA: Advanced Knowledge International), 19–42.

Di Paolo, E. A., Buhrmann, T., and Barandiaran, X. (2017). *Sensorimotor Life: An Enactive Proposal, 1st Edn*. Oxford: Oxford University Press.

Egbert, M. D., and Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Front. Hum. Neurosci.* 8, 590. doi: 10.3389/fnhum.2014.00590

Egbert, M. D., and Barandiaran, X. E. (2015). Corrigendum: Modeling habits as self sustaining patterns of sensorimotor behavior. *Front. Hum. Neurosci.* 9, 209. doi: 10.3389/fnhum.2015.00209

Egbert, M. D., and Cañamero, L. (2014). "Habit-based regulation of essential variables," in *Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, eds H. Sayama, J. Rieffel, S. Risi, R. Doursat, and H. Lipson (New York, NY: MIT Press), 168–175.

Egbert, M. D. (2018). "Investigations of an adaptive and autonomous sensorimotor individual," in *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, eds T. Ikegami, N. Virgo, O. Witkowski, M. Oka, R. Suzuki, and H. Iizuka (New York, NY: MIT Press), 343–350.

Fodor, J. A. (1968). *Psychological Explanation; An Introduction to the Philosophy of Psychology*. New York, NY: Random House Inc.

Gánti, T. (1975). Organization of chemical reactions into dividing and metabolizing units: the chemotons. *Biosystems* 7, 15–21. doi: 10.1016/0303-2647(75)90038-6

Graves, A., and Schmidhuber, J. (2009). "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems, Vol. 21*, eds D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (New York, NY: Curran Associates, Inc.)

Grey Walter, W. (1950). An imitation of life. *Sci. Am.* 182, 42–45. doi: 10.1038/scientificamerican0550-42

Hambrick, D. Z., Burgoyne, A. P., and Altmann, E. M. (2020). "Problem-solving and intelligence," in *The Cambridge Handbook of Intelligence, Cambridge Handbooks in Psychology, 2nd Edn*, ed R. J. Sternberg (Cambridge: Cambridge University Press), 553–579.

Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. (1997). Evolutionary robotics: the sussex approach. *Rob. Auton. Syst.* 20, 205–224. doi: 10.1016/S0921-8890(96)00067-X

Jacquey, L., Baldassarre, G., Santucci, V. G., and O'Regan, J. K. (2019). Sensorimotor contingencies as a key drive of development: from babies to robots. *Front. Neurorobot.* 13, 98. doi: 10.3389/fnbot.2019.00098

Luisi, P. L. (2006). "The emergence of life," in *From Chemical Origins to Synthetic Biology* (Cambridge: Cambridge University Press).

Maes, P. (1990). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. Cambridge, MA: MIT Press.

Man, K., and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* 1, 446–452. doi: 10.1038/s42256-019-0103-7

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Springer.

Maye, A., and Engel, A. K. (2013). Extending sensorimotor contingency theory: prediction, planning, and action generation. *Adapt. Behav.* 21, 423–436. doi: 10.1177/1059712313497975

MLPerf (2021). *MLCommons*. Available online at: https://mlcommons.org/

Newell, A., Shaw, J. C., and Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychol. Rev.* 65, 151–166. doi: 10.1037/h0048495

Newell, A., and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM* 19, 113–126. doi: 10.1145/360018.360022

Newman, S. A. (2018). "Inherency," in *Evolutionary Developmental Biology*, eds L. Nuno de la Rosa and G. Müller (Cham: Springer International Publishing), 1–12.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: The MIT Press.

Nolfi, S., Bongard, J., Husbands, P., and Floreano, D. (2016). "Evolutionary robotics," in *Springer Handbook of Robotics* (Berlin: Springer), 2035–2068.

Norvig, P., and Russell, S. (2021). *Artificial Intelligence: A Modern Approach, Global Edition*. Harlow: Pearson.

O'Regan, J. K., and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–973; discussion 973-1031. doi: 10.1017/S0140525X01000115

Pfeifer, R., and Scheier, C. (2001). *Understanding Intelligence*. Cambridge, MA: MIT Press.

Pinker, S. (1997). *How the Mind Works*. New York, NY: W. W. Norton & Company.

Putnam, H. (1965). "The mental life of some machines," in *Philosophical Papers: Volume 2, Mind, Language and Reality* (Cambridge: Cambridge University Press), 408–428.

Rosen, R. (1991). *Life Itself, 1st Edn*. New York, NY: Columbia University Press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386. doi: 10.1037/h0042519

Ruiz-Mirazo, K., Pereto, J., and Moreno, A. (2004). A universal definition of life: autonomy and open-ended evolution. *Origins Life Evolut. Biospheres* 34, 323–346. doi: 10.1023/B:ORIG.0000016440.53346.dc

Rumelhart, D. E., McClelland, J. L., and Group, P. R. (1988). *Parallel Distributed Processing, Vol. 1*. Cambridge, MA: MIT Press.

Simon, J. R., and Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics* 6, 99–105. doi: 10.1080/00140136308930679

Smithers, T. (1997). Autonomy in robots and other agents. *Brain Cogn.* 34, 88–106. doi: 10.1006/brcg.1997.0908

Steels, L. (1990). "Towards a theory of emergent functionality," in *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats* (Cambridge, MA: MIT Press), 451–461.

Varela, F. J. (1992). "Whence perceptual meaning? A cartography of current ideas," in *Understanding Origins: Contemporary Views on the Origin of Life, Mind and Society, Boston Studies in the Philosophy and History of Science 130, 1st Edn*, eds J.-P. Dupuy and F. J. Varela (Dordrecht: Springer Netherlands), 235–263.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind : Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Varela, F., Maturana, H., and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems* (Cambridge, MA). 5, 187–196. doi: 10.1016/0303-2647(74)90031-8

Vernon, D. (2010). Enaction as a conceptual framework for developmental cognitive robotics. *Paladyn J. Behav. Rob.* 1, 89–98. doi: 10.2478/s13230-010-0016-y

Webb, B. (1995). Using robots to model animals: a cricket test. *Rob. Auton. Syst.* 16, 117–134. doi: 10.1016/0921-8890(95)00044-5

Winograd, T., and Flores, F. (1987). *Understanding Computers and Cognition*. Boston, MA: Addison-Wesley Professional.

Zarco, M., and Egbert, M. (2019). "Different forms of random motor activity scaffold the formation of different habits in a simulated robot," in *Artificial Life Conference Proceedings* (New York, NY: MIT Press).

Ziemke, T., and Lowe, R. (2009). On the role of emotion in embodied cognitive architectures: from organisms to robots. *Cogn. Comput.* 1, 104–117. doi: 10.1007/s12559-009-9012-0

# Appendix A

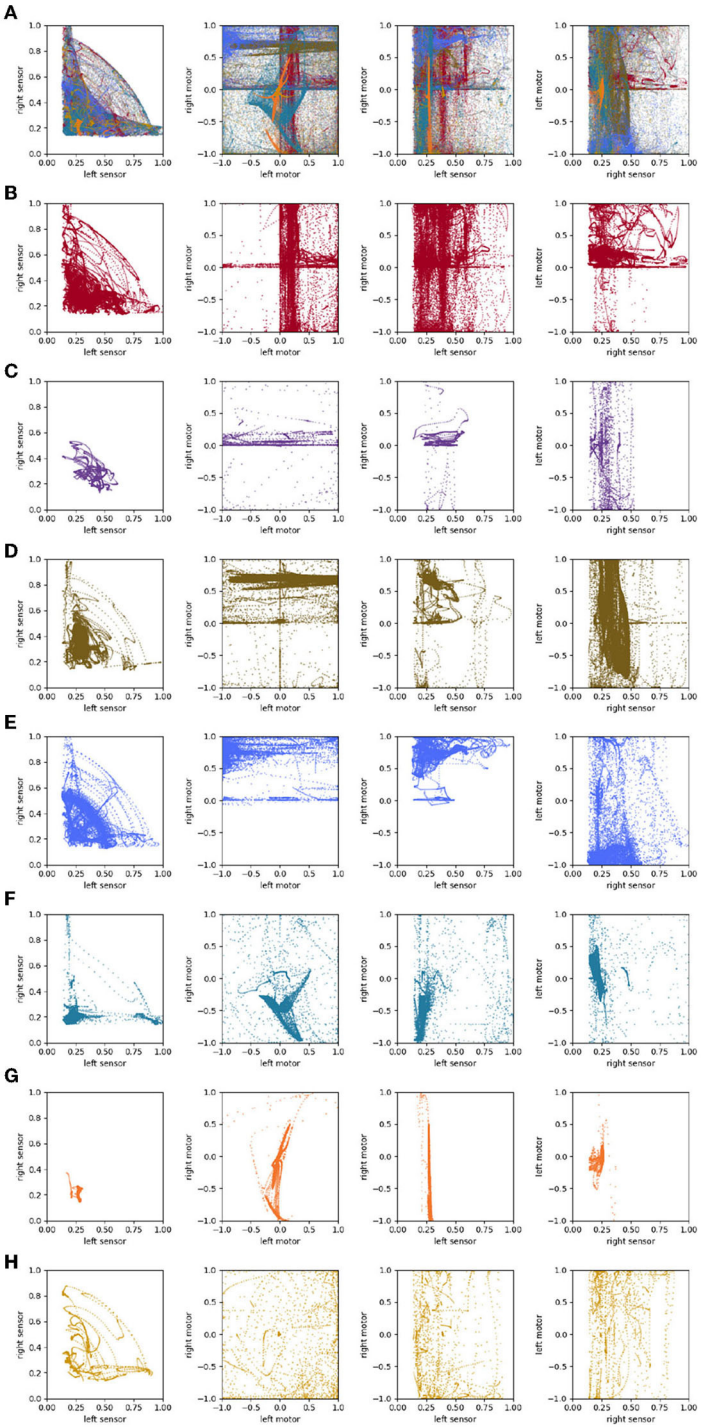



FIGURE A1
Projections of the sensorimotor trajectories of the visual robot. Colors and letters **(B–H)** are included to facilitate comparison with Figure 3. **(A)** Shows the superposition of trials **(B–H)**.

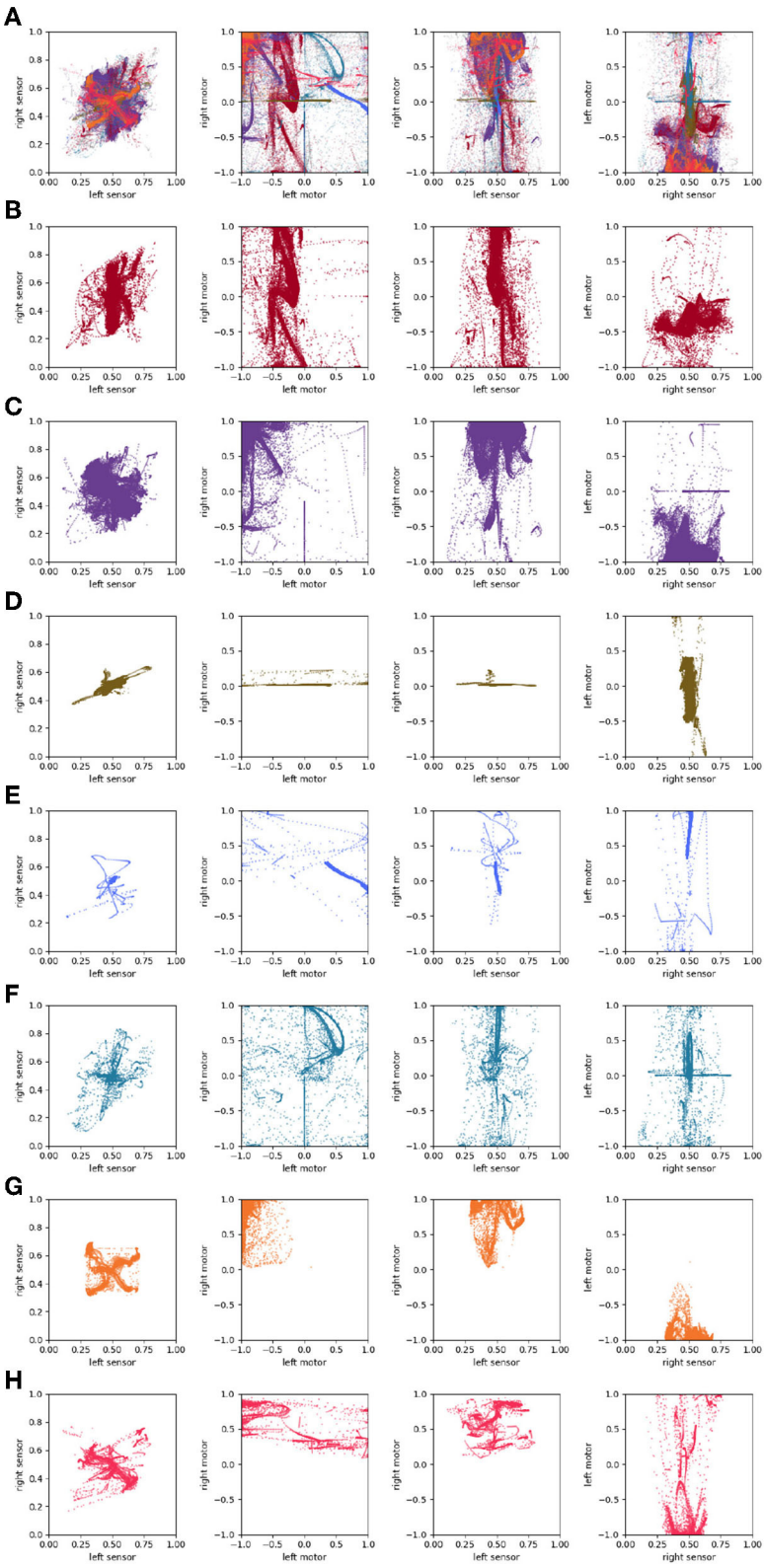



FIGURE A2
Projections of the sensorimotor trajectories of the visual robot. Colors and letters **(B–H)** are included to facilitate comparison with Figure 7. **(A)** Shows the superposition of trials **(B–H)**.

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fnbot.2022.944986

Check for updates

OPEN ACCESS

EDITED BY
Adam Safron,
Johns Hopkins Medicine, United States

REVIEWED BY
Vasileios Basios,
Université Libre de Bruxelles, Belgium
Alejandra Ciria,
Faculty of Psychology, National Autonomous University of Mexico, Mexico

*CORRESPONDENCE
Natalie Kastel
Natalie.Kastel@gmail.com

†These authors share first authorship

RECEIVED 16 May 2022
ACCEPTED 30 November 2022
PUBLISHED 09 January 2023

CITATION
Kastel N, Hesp C, Ridderinkhof KR and Friston KJ (2023) Small steps for mankind: Modeling the emergence of cumulative culture from joint active inference communication. *Front. Neurorobot.* 16:944986. doi: 10.3389/fnbot.2022.944986

COPYRIGHT
© 2023 Kastel, Hesp, Ridderinkhof and Friston. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Small steps for mankind: Modeling the emergence of cumulative culture from joint active inference communication

Natalie Kastel[1,2,3*†], Casper Hesp[1,2,4,5†], K. Richard Ridderinkhof[1,5] and Karl J. Friston[4]

[1]Amsterdam Brain and Cognition Centre, University of Amsterdam, Amsterdam, Netherlands, [2]Institute for Advanced Study, University of Amsterdam, Amsterdam, Netherlands, [3]Precision Psychiatry and Social Physiology Laboratory, Department of Psychiatry, CHU Sainte-Justine Research Center, Université de Montréal, Montreal, QC, Canada, [4]Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom, [5]Department of Developmental Psychology, University of Amsterdam, Amsterdam, Netherlands

Although the increase in the use of dynamical modeling in the literature on cultural evolution makes current models more mathematically sophisticated, these models have yet to be tested or validated. This paper provides a testable deep active inference formulation of social behavior and accompanying simulations of cumulative culture in two steps: First, we cast cultural transmission as a bi-directional process of communication that induces a generalized synchrony (operationalized as a particular convergence) between the belief states of interlocutors. Second, we cast social or cultural exchange as a process of active inference by equipping agents with the choice of who to engage in communication with. This induces trade-offs between confirmation of current beliefs and exploration of the social environment. We find that cumulative culture emerges from belief updating (i.e., active inference and learning) in the form of a joint minimization of uncertainty. The emergent cultural equilibria are characterized by a segregation into groups, whose belief systems are actively sustained by selective, uncertainty minimizing, dyadic exchanges. The nature of these equilibria depends sensitively on the precision afforded by various probabilistic mappings in each individual's generative model of their encultured niche.

KEYWORDS

**active inference, generalized synchrony, communication, social dynamics, cumulative culture, complex systems**

## 1. Introduction

The study of cultural evolution examines how processes of transmission and selection at the individual level bring about population level patterns of cultural change. As a general overarching trend, models of cultural evolution have seen a steady increase in complexity, resulting from specialized theories from social psychology on the

interconnected dynamics of culture. For instance, a recent model of cultural systems (Jansson et al., 2021) applied a framework that implements the structural dependencies between cultural traits and the emergent ways that these dependencies influence the acquisition of cultural transmission.

Another step in the direction of increasing the complexity and systems view of culture has been the investigation of the relationship between population structure and the capacity for a culture to accumulate beneficial cultural traits over time (i.e., cultural accumulation), which has been a particular topic of interest in the past decade. Although some empirical tests provide support for the hypothesis that effective population size constraints cumulative cultural evolution (Derex and Mesoudi, 2020), there is contradictory evidence regarding the relationship between population size and cultural accumulation (Kempe and Mesoudi, 2014). Another study theorizing about the foundations for the uniquely human capacity for cultural accumulation suggests that this capacity is rooted in a unique foraging niche which can still be observed in few hunter-gatherer populations. This niche, encompassing social interactions such as cooperation with unrelated individuals and social division of labor, underlies a task specialization which spreads cultural knowledge across individuals. This task division, a multilevel social structure which evolved as an adaptation to the environment, may explain human collective intelligence and its unique capacity for sophisticated cumulative culture (Migliano and Vinicius, 2022).

While the increase in the use of dynamical modeling in the literature on cultural evolution makes current models more mathematically sophisticated, these models have yet to be tested or validated (Kashima et al., 2017). Review of the literature reveals that one line of research that has been especially fruitful in that it can be validated experimentally are Bayesian models that create detailed models of cognition and have had remarkable success in producing predictions qualitatively in accord with experimental results (Kempe and Mesoudi, 2014). Currently these models have only been applied to relatively low level cognitive processes, but the creation of high level cognitive maps at the individual level as well as modeling the emergence of cultural change on the population-wide dynamics represents a promising line of future work.

The dynamics underlying the evolution of culture consist of three processes that are typically studied separately: the introduction of novel beliefs and practices to a culture (i.e., innovation), the transmission of established beliefs and practices within a population (i.e., innovation diffusion), and changes in their prevalence (Kashima et al., 2019). The term "cultural transmission" typically denotes the transference and spread of any particular fashion, ideology, preference, language, or behavior within a culture (Creanza et al., 2017). A prominent stream of quantitative models for cultural transmission are inspired by epidemiology, and convert models used for predicting the spread of a virus to formalize the spread of an idea (Bettencourt et al., 2006).

While the comparison of an idea to a (non-mutating) virus has its benefits from a formal perspective, it implies the controversial notion that an idea is simply copied during its transmission through cultural exchange between individuals. This notion is not only intuitively insufficient for a realistic characterization of communication dynamics, but also conflicts with established theoretical models of transmission on these same grounds.

Current literature in cultural psychology indicates that rather than being simply duplicated during transmission, cultural beliefs and practices are modified through the active interpretation of each individual (Kashima et al., 2019). Furthermore, psychological research indicates that effective human communication can be characterized by (usually only partially) common realities in which conversation partners share an intersubjective reference frame (Clark and Brennan, 1991; Echterhoff et al., 2009). Accordingly, conversation partners have to be understood as active participants that co-create these partially shared reference frames in a self-organizing fashion over the course of each interaction. Intersubjective theories of communication aim to account for those underlying dynamics that—slightly paradoxically—both enable and (to some extent) require the co-creation of (partially) shared reference frames amongst interlocutors. In contrast, traditional formulations have tended to oversimplify communication in terms of back-and-forth exchanges based on (largely) fixed symbolic meaning systems, implicitly presupposing those shared reference frames in an *ad-hoc* manner.

The theory of cumulative culture (Stout and Hecht, 2017; Dunstone and Caldwell, 2018) expands on the notion that progressive alterations of cultural beliefs and practices are intrinsic to all cultural exchanges because they are embodied, expressed, and interpreted differently by each individual participant of the ensemble (Dean et al., 2014). While efficient cultural exchanges do tend to be grounded in similar physical substrates across and within individuals (e.g., facial expressions, body language, etc.), the high abstraction levels and malleability of these substrates render cultural dynamics in a different class than phenomena that are wholly dependent on consistency across genetic substrates, such as sexual reproduction and disease spread. Of course, genetics does play a crucial role in the range of phenomena associated with gene-culture co-evolution, but that will be explored in future work.

Although there appears to be some degree of consensus on the intrinsic complexity of culture, it remains an outstanding challenge to provide convincing quantitative accounts of its full glory (Buskell et al., 2019). This article aims to act as a stepping stone toward tackling this challenge of characterizing cumulative culture by developing a multi-agent model based on deep active inference account. This work was developed by reformulating and greatly expanding upon a much shorter

conference contribution that was previously published by the two lead authors (Kastel and Hesp, 2021) and publicly available as an open-source preprint. The three Figures that were adapted or reprinted from that conference paper have been highlighted (Figures 2, 8, 9). Textual overlap has been minimized and reprinted Figures have been highlighted where relevant.

## 2. Methods

An emerging conclusion from the literature is that the term "transmission" for describing the spread of cultural information seems impoverished, as it leaves out the retention of cultural information. As implied by active inference—and theoretical models of communication—the acquisition of cultural beliefs is as fundamental to the understanding of cultural information spread as their transmission. For this reason, we will henceforth be referring to what is known in the literature on cultural transmission as communication, or more technically: the local dynamics of cumulative culture. Although we will use the term "communication" and "transmission" interchangeably in this paper, It is important to note that communication does not always imply cultural transmission. Cultural transmission, also known as cultural learning, refers to the learning of social behaviors that occurs in every new generation in a particular society (Nicol, 1995). Since cultural transmission only occurs when social behaviors or beliefs are learned, communication only truly implies transmission when what is being communicated has been picked up and solidified in the receiver's cognitive model. We assume this kind of communication in our simulations, which is why under our account communication does imply transmission.

Conversely, cultural transmission can occur without communicating information through language. It has been suggested that humans learn the social behaviors of their culture through immersive participation in cultural practices that selectively shape attention and behavior. This is a form of implicit learning, where agents infer other agents' expectations about the world and how to behave in a social context. It is even argued that implicit learning of information about other people's expectations constitutes the primary domain of statistical regularities that humans leverage to predict and organize behavior (Veissière et al., 2020). Although cultural transmission does not necessarily require verbal communication, we assume this kind of communication in our simulations.

### 2.1. Simulating the local dynamics of communication

In our model, cultural transmission is cast as the mutual attunement of actively inferring agents to each other's internal belief states. This builds on a recent formalization of communication as active inference (Friston and Frith, 2015) which resolves the problem of hermeneutics, (i.e., provides a model for the way in which people are able to understand each other precisely, despite lacking direct access to each other's internal representations of meaning) by appealing to the notion of generalized synchrony as signaling the emergence of a shared narrative to which both interlocutors refer. From the perspective of active inference, agents of a socio-cultural system infer the belief states of those in their environment and update their own representations accordingly. An emergent property of this bi-directional inference—and implicit belief updating—is the synchronization of belief states among the cultural ensemble (Palacios et al., 2019).

In nature, generalized synchrony emerges from a specific coupling between the internal states of dissipative chaotic systems (Pikovsky et al., 2003). As a fundamental concept in complex systems theory closely related to stochastic resonance (Nicolis and Nicolis, 2012), it is typical of complex nonlinear dynamics that afford the coexistence of chaotic and ordered subsystems (also called chimera patterns; see, e.g., Zakharova, 2020; Haugland, 2021). In active inference, the coupling of agents' internal states is made possible through communication, as it allows interlocutors to mutually influence each other and enter into a bidirectional action-perception cycle that can be described as coupled dynamical systems (Friston and Frith, 2015; Constant et al., 2018). When active inference agents engage in the coupled dynamics of communication, generalized synchrony between their internal states emerges from their mutual efforts to minimize uncertainty—as scored mathematically with (variational) free energy. Put simply, generalized synchrony ensures the greatest mutual predictability error resolves the greatest amount of collective uncertainty.

Our model of communication builds on the notion of generalized synchrony to suggest that the emergence of synchrony from the coupled communication of active inference agents may be operationalized as a particular convergence between their respective generative models. That is, when we simulate the belief-updating dynamics of communicating agents, the cultural reproduction of a particular idea takes the form of a learnable convergence between their respective belief states (expressed as generative models) and distinct representations combine into one synchronized, shared model of the world.

Formally, our model defines perceptual inference in terms of a coupling parameter linking the internal states of interlocutors through dialogue (Figure 1). Also understood as sensitivity to model evidence (A1), perceptual inference is a direct and explicit form of coupling that occurs over the span of a single dialogue such that it is hypothesized to modulate agents' convergence of internal belief states during communication.

While throughout the narrative of this paper we have characterized the alternative "idea" as having actual content, we

have intentionally left it unspecified such that it could also be taken to simply refer to a blanket disagreement with the ideas or practices representing the "status quo." In that sense, it is consistent with simulation work on social dynamics suggesting that cultural extremism can arise without the formulation of alternatives (e.g., see Kashima et al., 2021).

## 2.2. Simulating the global dynamics of cumulative culture

Cultural beliefs and practices spread within a society through communication, a process which we have referred to as the local dynamics of cumulative culture. This description is appropriate because the accumulated outcomes of each (local) dyadic interaction collectively determine the degree to which an idea is prevalent in a culture. Moving from local communication dynamics to the prevalence of a communicable idea—in a cumulative culture—is what we will refer to as the global dynamics of cumulative culture.

In our simulations of a cumulative culture, 50 active inference agents simultaneously engage in local dyadic communication as illustrated. Forty-nine of our 50 agents were initialized as adhering to a similar idea, which could be regarded as the status quo (indicated with the color red later on), while the initial strength of their adherence to this consensus varied across individuals because we generated the parameters of their generative models from various probability distributions to characterize variability in the population (described in the Appendix). The same holds for other modulators of cognitive, affective, and behavioral variability, such as (1) expectations about each other's expressions, (2) habit formation, and (3) emotional valence states (all described below). Jointly, the emergent effects of these individual differences gave rise to factions that vary in their adherence to the current consensus, in a way reminiscent of political diversity in real-life cultural environments: strict conservatives, centrists, and skeptics (see also Figure 6 below). In order to illustrate this spectrum, we introduced one agent (labeled "rogue agent" in Figure 6) whose idea strongly contradicted the consensus and who was fully resistant to the consensus idea. When we introduced this agent adhering to a divergent idea to the population, it propagated *via* pseudo-random engagements of agents in dialogue. In this simulated world of actively inferring agents, their individual mental (generative) models were slightly modified with every interlocutor they encounter, as their distinct representations converged to a shared narrative (Friston and Frith, 2015; Constant et al., 2019). The attunement of interlocutors to each other's generative models on the (local) microscale thus translated over time and with multiple encounters into collective free energy minimization on the (global) macroscale.

# 3. A generative model of communication

For a formal (variational free energy) proof of principle, we offer an active inference account of cultural dynamics. A foundational step in this endeavor is the formulation of generative models underlying the decision making of agents that can be deployed in simulations.

Active inference assumes that the brain-body system mimics a Bayesian inference machine by embodying a model of itself acting in the world and using local observations to secure evidence for that model. This model is "generative" in that it generates predictions of what observational data should look like, given that the model is correct. Predictions are compared with actual observations and any discrepancies (known as "prediction errors") are used to update the generative model (Smith et al., 2019). This (Bayesian belief) updating can be at a fast timescale corresponding to inference about the hidden causes of observations—or at a slow timescale corresponding to learning the parameters of the generative model, which best explain the inferred causes. For an elaboration on the mathematical foundations of active inference, the reader is referred to Friston K. et al. (2017).

In our simulations, agents attempt to convince each other of a cultural belief under generative models that operate with local information only. We formulate these generative models as a partially observed Markov decision process (MDP), where beliefs take the form of discrete probability distributions (for technical details on MDPs in active inference, see Hesp et al., 2019). To simulate active inference under these models, one specifies variables—such as hidden states ($x$, $s$), observable outcomes ($o$) and one-step action policies ($u$)—alongside parameters specifying the probabilistic relationships between the variables in question.

Agents' recollection of a visit is thus an expression of humans' innate ability to infer each other's expectations, which makes human cognition, sociality, and culture possible at all (Veissière et al., 2020). This rests on the idea that humans, having evolved to rely on elaborate and highly coordinated action, have expectations regarding other agents' sharing aspects of their own generative model, and thereby believing that other agents have those expectations as well. These carefully and implicitly coordinated and co-constructed expectations allow agents from a particular culture to learn what to expect from each other and leverage those expectations to act accordingly in their environment. In our model these expectations are manifested as agents' information and preference-seeking, which are biased toward the selection of similar interlocutors to engage with, in conversation.

MDPs allow for the construction of a deep hierarchical model comprising nested levels of complexity. Below we will

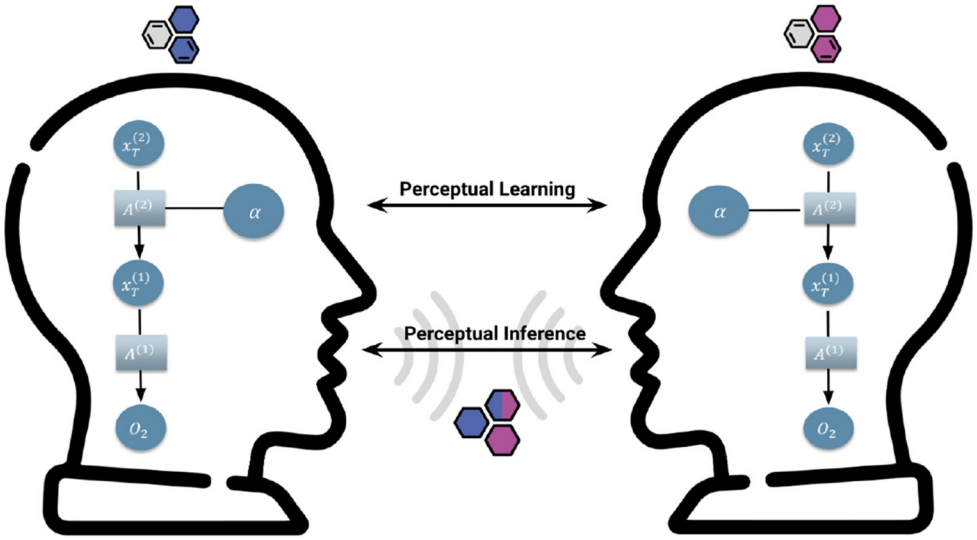



FIGURE 1
Communication Coupling Parameters. Our model defines two groups of parameters that couple the internal states of agents: Learning and inference. Perceptual learning (A2) is the learning of associations between emotional valence and belief states that guide the long term actions of our agents who hold and express beliefs. This learning happens at slow time scales, accumulating across multiple interactions and used to modify models over extended periods of exchange. Perceptual Inference (A1)—namely, sensitivity to model evidence—operates on fast time scales and is direct and explicit to agents during dialogue. Importantly, we hypothesized that without precise evidence accumulation, agents would be insensitive to evidence regarding the belief state of the other, and their internal states would not converge.

describe those levels and detail the cognitive processes that take place within each one (Figure 2).

For our simulations, six kinds of matrices were parameterized ($A$, $B$, $C$, $E$, $C$, and $G$) using two kinds of concentration parameters ($\alpha$,$\varepsilon$) for Dirichlet distributions, and temperature and rate parameters for precision terms (indicated with $\gamma$; see Figure 3 and Appendix A9).

## 3.1. Perceptual inference

The first level of this generative model captures how agents process belief claims they are introduced to through conversation with other agents. The perception of another's beliefs requires prior beliefs (represented as likelihood mapping $A1$) about how hidden states ($s1$) generate sensory outcomes ($o$). Specifically, our agents form expectations about the likelihood of encountering expressions of particular ideas, given their beliefs about the degree of consensus in the agent's social circles and their past experiences with individual agents. Parameterizing this (likelihood) mapping in terms of precision can be understood as parameterizing each agents' sensitivity to the claims of others. High precision here corresponds with high sensitivity to claims. The likelihood precisions for each agent were generated from a continuous gamma distribution, which was skewed in favor of high sensitivity to evidence at the population level (see Figure 2: Perception).

### 3.1.1. Perceptual inference as a coupling parameter

The sensitivity to another agent's claims ($A1$), represents the explicit coupling between interlocutors, in terms of how much belief updating one agent can induce in another agent. It is a key element in our simulations (Figure 1). We call this parameter explicit because it modulates the direct (i.e., explicitly articulated) and immediate (i.e., occurs over the course of a single interaction with an agent) influence of agents' claims on the beliefs of others (Friston and Frith, 2015). In other words, sensitivity to claims—encoded by the likelihood precision—couples the belief states of interlocutors *via* their claims or utterances to each other (Figure 8). Crucially, belief updating depends not only on their adherence to each other's claims but also a certain (varying) degree of commitment to their own beliefs. The balance is determined by each agents' sensitivity to sensory evidence; i.e., the claims of interlocutors.

Technically, we can describe belief updating in terms of the generative model in Figure 3 as follows:

- Initial higher-level core support for the idea at the beginning of the simulation ($T = 1$):

$$P(x^{(2)}_{core,T=1}) = \mathrm{Cat}(\mathbf{D}^{(2)}) \quad (1)$$

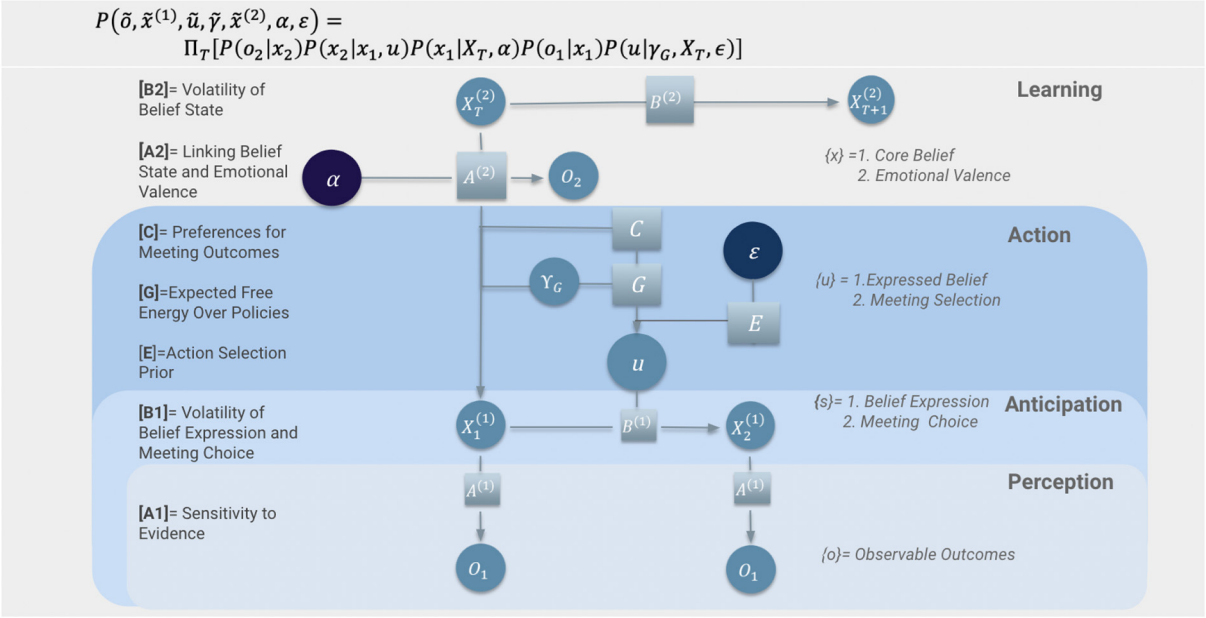



FIGURE 2
A generative model of communication. This Figure was reprinted from an open-source preprint of a conference paper, with permission of the authors (Figure 1 of Kastel and Hesp, 2021). Variables are depicted as circles, parameters as squares and concentration parameters as dark blue circles. Visualized on a horizontal line from left to right, states evolve in time. Visualized on a vertical line from bottom to top, parameters underwrite a hierarchical structure that corresponds to levels of cognitive processing. Parameters are listed on the left of the generative model and variables are on the right.

- Evolving higher-level beliefs after each meeting ($T > 1$), introducing volatility over time:

$$P(x^{(2)}_{core,T}|x^{(2)}_{core,T-1}) = \text{Cat}(\mathbf{B}^{(2)}) \quad (2)$$

- Initializing lower-level beliefs about the claims of others, based on higher-level (cross-meeting) beliefs:

$$P(x^{(1)}_{idea}) = \text{Cat}(\mathbf{x}^{(2)}_{core,T}) \quad (3)$$

- Updating beliefs about the other agent's belief based on their claims (Appendix A7), within the current meeting:

$$Q(x^{(2)}_{idea}) = Q(x^{(1)}_{idea}) = \text{Cat}(\mathbf{o}_{expr}) \quad (4)$$

- Updating of core belief based on claims of self and another agent after each meeting (detailed descriptions of the computations involved in this belief updating can be found in the Appendix):

$$Q(x^{(2)}_{core}) = \sigma(\ln \mathbf{x}^{(2)}_{core} + \gamma^{(2)}_{A,self} \ln \mathbf{o}_{expr,self} + \gamma^{(2)}_{A,other} \ln \mathbf{o}_{expr,other}) \quad (5)$$

## 3.2. Anticipation

At the first level, our generative model specifies the agents' beliefs about how hidden states (detailed in Appendix A2) evolve over time. The precision of state transition probabilities in B1 (Figure 3C) specifies the volatility of an agent's meeting location ($s2$) and beliefs in particular claims ($s1$) [$B1$]. For each agent, this precision parameter is sampled from a gamma distribution, determining the a priori probability of changing state, relative to maintaining a current state. Note that belief states themselves are defined on the continuous range $(0, 1)$ (i.e., as a probability distribution on a binary state), such that repeated state transitions tends to result in a continuous decay of confidence over time, in the absence of new evidence (where the rate of decay is inversely proportional to the precision of $B1$) (see Figure 2: Anticipation).

## 3.3. Action

After inferring and anticipating hidden states, our agents conduct deliberate actions to minimize expected free energy (the generative model for action is detailed in Appendixes A4, A5). At each time point, a policy ($u$) is chosen out of a set of possible action sequences. In our simulations, two types of actions are allowed: selecting an agent to meet at each given time point ($u2$) and selecting a specific claim to express in conversation ($u1$). The first allowable action covers 50 possible outcomes (one for each agent in the simulation) while the

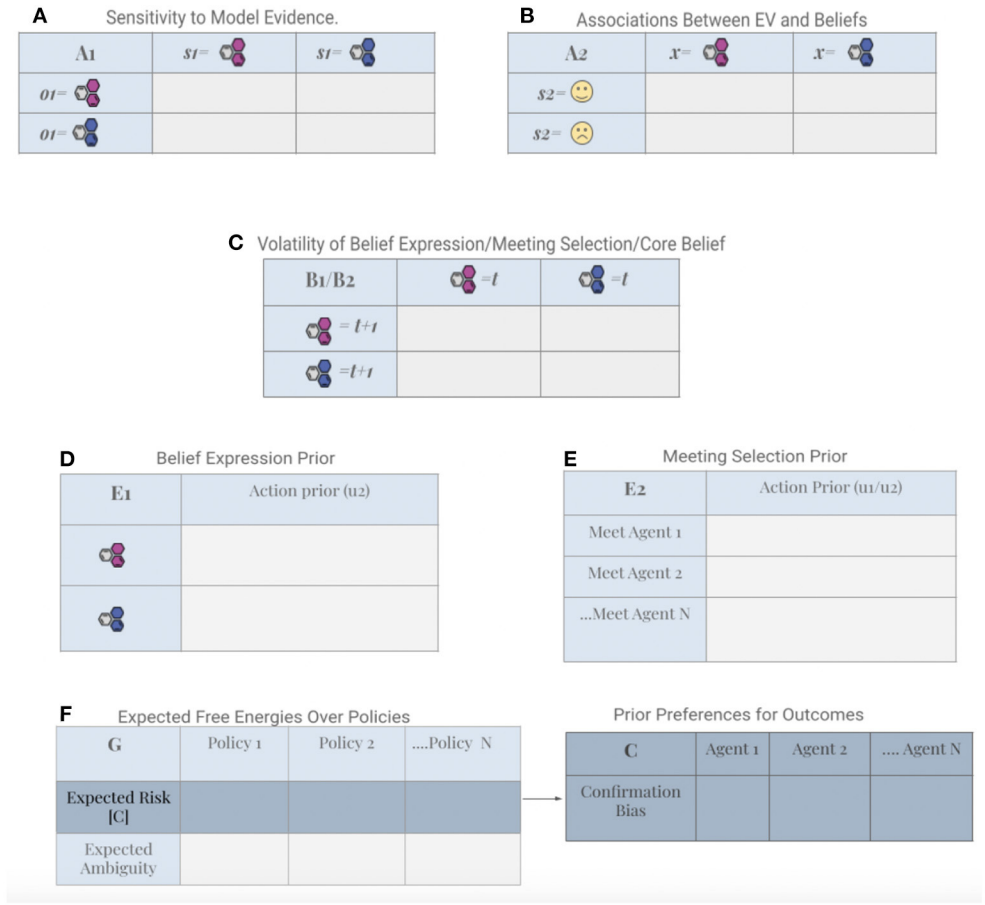



FIGURE 3
Generative model parameters. **(A)** The *A*1 matrix specifies an agents' perception of an interlocutors' expressed beliefs. The precision of this likelihood mapping determines the agent's sensitivity to these expressions. **(B)** The *A*2 matrix represents what the agent has learned about the mapping between her high and low level beliefs. **(C)** The *B* matrix, or state transition probabilities, represent what the agent has learned about how hidden states evolve over time. The precision of B matrices can be understood as encoding the volatility of belief states. **(D)** The *E*1 matrix is one of two habitual contributions to action selection. It covers two possible outcomes for expressing beliefs. This contribution is specified on a continuous range between (0,1), where the extremes correspond to either complete confidence in denying or supporting the claim. **(E)** The *E*2 matrix is the second habitual component for action, and it holds 50 possible outcomes for meeting selection (i.e., the probability for meeting each agent in the population). **(F)** The expected free energy of allowable policies (i.e., choices or actions) is indicated with *G*, which entails two components: 1. expected risk (the KL-divergence from the *C* matrix and biasing toward confirming one's preferred ideas) and expected ambiguity, which biases toward meeting new agents with unknown beliefs. Note: The purpose of this Figure is to draw the attention of the reader to the general form of the matrices shared across the simulated agents. The tables are left empty because, for any individual agent in the simulated population, each of these objects contains specific numbers, which are initially generated procedurally from various probability distributions (described in the text) and change throughout the simulation as the agents interact in their shared environment. Specific numbers could at best describe only one particular agent at a given instance of time (which does not represent the entire population). Furthermore, the probability distributions used to generate initial values do not reflect the additional steps required such as, e.g., the renormalization procedures involved in applying a softmax operator. Finally, it would also occlude the fact that certain entries (e.g., the expected free energy) will vary over time during a simulation.

second corresponds to denying or supporting the claim. In order to simulate variability in an agents' confidence in a belief claim, the claim is generated for each conversation from a beta distribution that is parameterized by the speaker's (phenotype-congruent) action model. Each policy under the *G* matrix (Figure 3F) specifies a particular combination of actions, and the policy that minimizes expected free energy is chosen (see Figure 2: Action).

### 3.3.1. Habitual belief expression and meeting selection

At the low level of cognitive control, each agent starts with a baseline prior expectation concerning the probability of a particular policy being selected (action prior probabilities [*E*1 and *E*2], Figures 3D,E). This parameter can be understood as modeling a habitual cognitive process, where an agent's current one-step policy ($u$) is biased toward previously selected actions

($u1$, $u2$). In our simulations, agents observe and track previous actions *via* the accumulation of a concentration parameter ($\varepsilon$), thus enabling continued updates to action priors, which can either strengthen or weaken previous habits (details for habit learning of belief expression and meeting selection are provided in Appendix A8).

### 3.3.2. Voluntary belief expression

At the high level of cognitive control, agents incorporate a series of processes underlying the selection of a particular claim for expression ($u2$). In addition to habitual factor (E), this selection involves several considerations. First, an agent considers their core belief state ($x$), and the way this state a priori maps on to one of two discrete emotional valence states ($s2$) *via* a likelihood mapping [$A2$] (Figure 3B). Emotional Valence ($EV$) is defined as the extent to which an emotion is positive or negative (Russell and Barrett, 1999), such that agents' core beliefs are a priori associated with either positive emotional valence or negative emotional valence (with some probability). As a minimal form of vicarious learning, the initial mapping is further updated based on associations agents observe between their interlocutors' expressed claims and $EV$-value (details of the generative process underlying belief expression and emotional valence are provided in Appendix A6). The initial mapping therefore involves minimal precision for the expected $EV$ for the alternative belief since agents are first introduced to this belief (and associated $EV$) during the simulations. For this reason, the initial likelihood mapping between states is updated throughout our simulation *via* a crucial concentration parameter ($\alpha$) which will be elaborated on under level 4.

The inferred $EV$ state is then used to generate an action precision ($\gamma$) such that positive $EV$ generates high confidence in action selection ($u1$) and negative $EV$ generates low confidence. Higher confidence values produce higher precision on the expected free energy ($G$) for one's belief claim expressed in the current conversation.

EV states are generated from core belief states, using a (learnable) likelihood mapping:

$$P\left(x^{(1)}_{sat}|x^{(2)}_{core}\right) = Cat\left(\mathbf{A}^{(2)}_{sat}\right) \tag{6}$$

Confidence of belief expression is generated using a Gamma distribution, where the rate parameter $\beta_{expr}$ is the Bayesian model average of $\beta^{(+,-)}$ values associated with high and low satisfaction:

$$P\left(\gamma_{expr}\right) \approx \Gamma\left(1, \boldsymbol{\beta}_{expr}\right) \tag{7}$$

$$\boldsymbol{\beta}_{expr} = \beta^{(+,-)} \cdot \mathbf{x}^{(1)}_{sat} \tag{8}$$

where $\beta^{(+,-)} = [0.25; 2.0]$

The expression of beliefs is guided by current core beliefs (scaled with satisfaction-dependent $\gamma_{expr}$) and by habitual belief expression $\mathbf{E}_{expr}$ (scaled with a fixed parameter $\gamma_{E,expr}$):

$$P\left(u_{expr}|\gamma_{expr}\right) = \sigma\left(-\gamma_{expr} ln\, \mathbf{x}^{(2)}_{core} + \gamma_{E,expr}\mathbf{E}_{expr}\right) \tag{9}$$

The intrinsically stochastic and itinerant nature of the generative process of communication is modeled by using a two-dimensional Dirichlet distribution to generate observed expressions on the range (0,1), where each agent's belief expression prior $P\left(u_{expr}|\gamma_{expr}\right)$ is used to specify their concentration parameters (multiplied by 12 to reduce variance):

$$\mathbf{o}_{expr} = \mathrm{Dir}(12\mathbf{u}_{expr}) \tag{10}$$

### 3.3.3. Voluntary meeting selection

While the choice of interlocutor is predetermined in a dyad, our multi-agent simulations required a specification of the process behind agents' selection of a conversational partner ($s3$) at each of the (hundred) time points. Building on previous work on active inference navigation and planning (Kaplan and Friston, 2018), meeting selection in our model is represented as a preferred location on a grid, where each cell on the grid represents an agent to meet (Figure 4).

Importantly, agents differ in their action model of which agent to visit at each time point. Their individual choices are guided by expected free energy G (Figure 3F) which entails maximizing the expected utility of an action (known as pragmatic value) as well as maximizing the expected information gain (known as epistemic value). These two values constrain each other such that maximizing both simultaneously is partially (but not entirely) paradoxical (as illustrated in Figure 4). These constraints may also be understood as formalizing the exploration-exploitation trade-off, where epistemic value (exploration) refers to the benefit of searching to get a better estimation of promising areas that offer pragmatic value (exploitation) (Friston and Frith, 2015).

Mathematically, action selection was formalized as follows:

$$P\left(u_{visit}|\gamma_{G,visit}\right) = \sigma\left(-\gamma_{G,visit}\mathbf{G}_{visit} + \gamma_{E,visit}\mathbf{E}_{visit}\right) \tag{11}$$

Here, $\mathbf{G}_{visit}$ represents the expected free energy:

$$\mathbf{G}_{visit} = \mathbf{o}_{expr,visit} \cdot \left(\ln \mathbf{o}_{expr,visit} - \mathbf{C}_{idea}\right) + \mathbf{H} \cdot \mathbf{x}_{idea,visit} \tag{12}$$

An agent traveling to visit:

$$x^{(1)}_{idea,visit} = \mathbf{B}^{(1)}_{visit} x^{(1)}_{idea,home} \tag{13}$$

Expectations about the support for an idea expressed by each potential agent one could visit:

$$\mathbf{o}_{expr,idea} = \mathbf{A}^{(1)}_{idea}\mathbf{x}^{(1)}_{idea,visit} \tag{14}$$

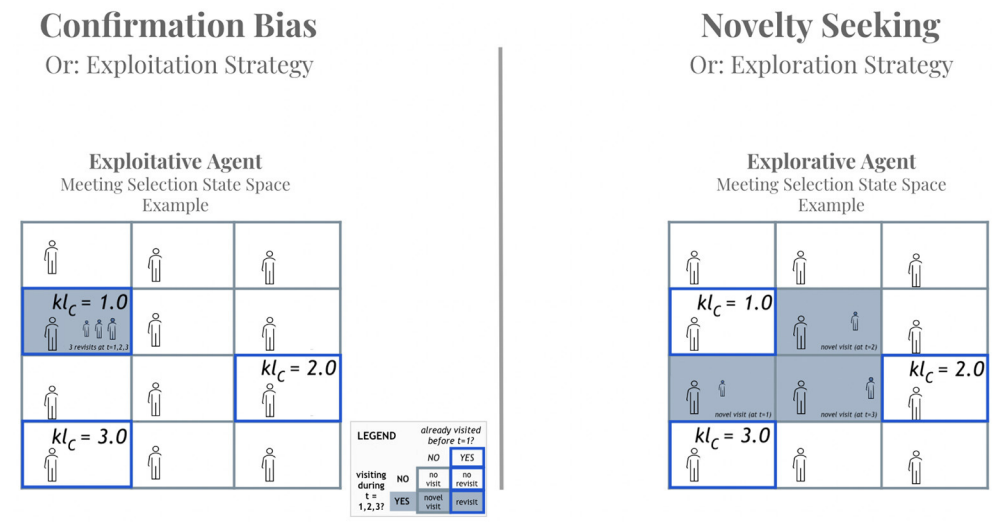



FIGURE 4
A simplified example of exploitation- and exploration-driven strategies for meeting selection. This Figure illustrates the behavioral differences between the extreme cases of being fully driven by exploitation **(Left)** or exploration **(Right)**. Each cell on the grid corresponds to a potential interlocutor for these agents, who make decisions in three consecutive time steps ($t = 1, 2, 3$) and have previously engaged with three other interlocutors (marked with blue rectangles), where we use the shorthand $kl_C$ to indicate the pragmatic component of the expected free energy $G_{pragmatic,visit} = o_{expr,visit} \cdot (\ln o_{expr,visit} - C_{idea})$, which corresponds to the KL divergence between expectations about the interlocutor at that location (as informed by previous visits) and the preferred ideas of our agents, such that lower values correspond to a better match. Cells that are visited during $t = 1, 2, 3$ are filled with granite. The exploitation-driven agent **(Left)** simply revisits three times a known interlocutor with the lowest $KL_C$. In contrast, the exploration-driven agent **(Right)** prefers novel visits and switches to an unknown agent every time step. In the simulations presented later, agents will dynamically balance these two strategies as their preferences themselves evolve over time.

Individual preferences about the support for the idea:

$$\mathbf{C}_{idea} = \ln\left(\mathbf{A}_C^{(2)}\mathbf{x}_{core}^{(2)}\right) \tag{15}$$

Finally, expectations about a potential reduction in ambiguity about the support for an idea by a particular agent reflects one's recollection about their most recent visit to this other agent. $H_j = 0$ if an agent can remember a recent visit (i.e., there is no ambiguity left to reduce), and 0.1 otherwise.

Crucially, both types of (information and goal seeking) preferences are absorbed into expected free energy. Pragmatic value translates into a bias toward meeting agents with similar beliefs at a given time point. This bias reflects the widely observed phenomenon in psychology research that people's choices tend to be biased toward confirming their current beliefs (Nickerson, 1998). Confirmation bias, or a state-dependent preference ($C$) for meeting "belief compatible" agents, biases action selection through the risk component of expected free energy ($G$) (Figure 3F). Under active inference, a preference for meeting agents with similar beliefs increases the propensity for generalized synchronization, which underwrites the emergence of (expected free energy reducing) shared expectations (Hesp et al., 2019).

In contrast, emphasizing epistemic value translates into a bias toward meeting agents whose beliefs are unknown at a given time point. This bias reflects the extent to which agents are driven by the minimization of the ambiguity component of expected free energy ($G$; Figure 3F) about the beliefs of other agents. Novelty seeking, or a proclivity for encountering novel agents with unknown beliefs is a strategy for maximizing information gain. Also understood as intrinsically motivated curiosity behavior (Friston K. J. et al., 2017), maximization of epistemic value helps individuals to better predict the consequences of their actions (e.g., when they decide which agent to meet) as they reduce uncertainty about hidden states of their environment, whether real or imagined (e.g., in this case it refers to the ideas supported by other agents).

Because of our method of procedural generation of various (hyper)parameters from various probability distributions (described in more detail in the Appendix), a continuous spectrum of cognitive, behavioral, and affective tendencies emerged in our simulations due to the large variety of possible combinations. In principle one could obtain any range of behaviors from this method of procedural generation based on a set of probability distributions for all these hyperparameters, which could hence be fitted to population data. A full-blown analysis of all the emergent variability in the simulated populations is beyond the scope of the current paper.

For the sake of our demonstration, a clear distinction between agents with high and low confirmation bias was introduced in our simulations by drawing individualized

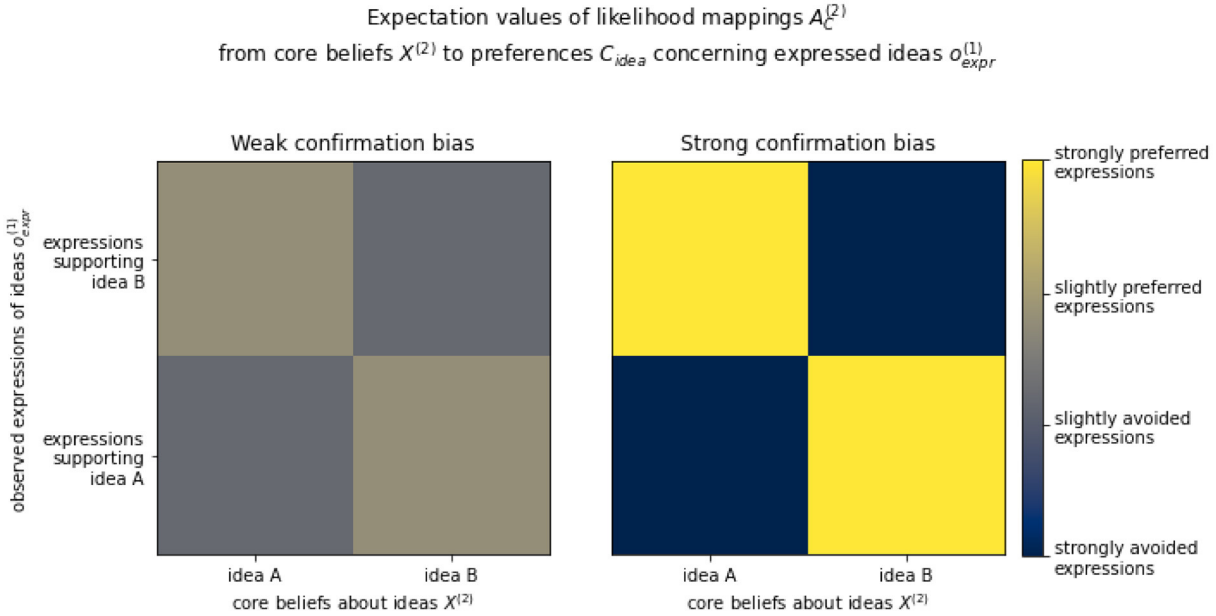



FIGURE 5
Two sets of expectation values of the Dirichlet distributions used to generate top-down likelihood mappings $A_C^{(2)}$, from core beliefs about ideas $X^{(2)}$ to preferences concerning expressed ideas $C_{idea}$, representing the two distinct populations for which parameters were initialized with different degrees of confirmation bias. Weak confirmation bias **(Left)** corresponded to mild preferences for observed expressions to confirm core beliefs, while strong confirmation bias **(Right)** corresponded to a strong preference for observed expressions to confirm core beliefs (essentially a one-to-one mapping).

hyperparameters from two distinct sets of Dirichlet distributions (illustrated in Figure 5; described in the Appendix) to obtain each agent's likelihood mapping from higher-level core beliefs to lower-level preferences concerning observed expressions. The resulting distinct populations could have emerged from, e.g., cultural segregation where different cultural subgroups have developed different priorities in guiding social interactions—in this case guided more or less strongly by confirmation of core beliefs.

Novelty-seeking tendencies were not explicitly coded and simply emerged from the parameters that regulate the relative impact of epistemic vs. pragmatic value in the expected free energy, although it should be clear that high confirmation bias tends to suppress novelty-seeking. In Figure 6, the distinction between "strict conservatives," "centrists," and "skeptics" was used to qualitatively describe the emergent continuous spectrum purely for communicative purposes and should not be taken as a definite discretization.

## 3.4. Perceptual learning

On this level of belief updating, agents learn contingencies; for example, how core belief states (specified in Appendix A1) change over time (*B*2) (Figure 3C). This is the highest level of cognitive processing, where agents learn (as detailed in Appendix A3). By talking with other synthetic agents and inferring their emotional and belief states, our agents learn associations between EV and beliefs *via* a high level likelihood mapping (*A*2), (updated *via* concentration parameter $\alpha$). The updating of the likelihood mapping between beliefs and claims, is detailed in Appendix A7. This kind of learning is important because it provides our agents with certainty, regarding the emotional value they can expect from holding the alternative belief to the status quo, which has low precision at the beginning of the simulation (before the population is introduced to an agent proclaiming this belief).

### 3.4.1. Perceptual learning as a coupling parameter

The learning of associations between belief and emotional valence states may be understood as a form of implicit coupling between agents (Figure 1), in that it represents an indirect and secondary influence of one agent's internal state on another. That is, sensitivity to each others' mental states is made possible only through inferences about the others' emotional state (in the absence of any overt or observable evidence for that emotional state).

In contrast to perceptual inference, learning occurs at slow time scales as mutual minimization of prediction error brings about a convergence in the parameters of hierarchical models

that generate mutually sympathetic (or possibly empathetic) predictions. Parameter learning accumulates across multiple interactions, modifying generative models over a long period of time as opposed to being immediately expressed in agents' behavior. This is why perceptual learning does not bring about an immediate convergence or synchrony between interlocutors' internal states, but is only expressed in agents' adapted behavior over time.

Individuals vary in the degree to which they are sensitive to the information gained by learning associations between belief states and their potential emotional outcomes. This variation is represented in each agent's categorical probability distribution $A2$ that is updated throughout the simulation *via* a concentration parameter ($\alpha$) as they accumulate information with every agent they meet. Updates to the $A2$-concentration parameters model the way in which agents' associations between belief and emotional states are based on implicit observations of others' emotional states.

The prior for this likelihood mapping is specified in terms of a Dirichlet distribution:

$$P\left(A_{sat}^{(2)}\right) = \mathrm{Dir}\left(\alpha_{sat}^{(2)}\right) \tag{16}$$

The associated approximate posterior accumulates the precision-weighted counts of correspondences between observed expressions and satisfaction levels:

$$Q\left(A_{sat}^{(2)}\right) = \mathrm{Dir}\left(\alpha_{sat}^{(2)} + \gamma_A^{(2)}\mathbf{o}_{expr}\ln \mathbf{x}_{sat}^{(1)}\right) \tag{17}$$

## 3.5. Simulated process summary for multi-agent social dynamics

For additional clarity, we provide a verbal and graphical summary of multi-agent social dynamics in our simulations. The "community square" (depicted in Figure 6) contained 50 social agents with deep generative models (identical structure + individualized parameters), simulating how they mingled (step 1) and conversed daily (step 2) about two mutually exclusive ideas ("red" and "blue"), as illustrated in Figure 7.

### 3.5.1. Simulated dynamics within days

Within each day, every agent engages in steps 1 and 2, generating:

- Expected support for idea (from self and others)
- Expected claim expressions (from self and others)
- Current claim preferences
- Current valence state
- Current action model precision
- Memory of most recent visit of other agents
- Current selection of an agent to visit

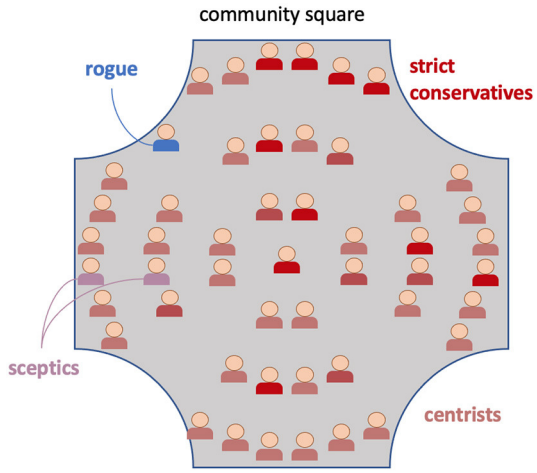



FIGURE 6
A depiction of the community square with its initial opinion distribution: One idea (red) was supported by almost all agents, with some variation due to individualized model parameters, roughly dividing into "strict conservatives," "centrists," and a few "skeptics." At first, the alternative idea (blue) is supported by only one, stubborn agent ("rogue").

- Expressed opinions (when visiting and when visited)
- Expressed affective cues (when visiting and when visited)

### 3.5.2. Simulated dynamics across days

Across days, every agent maintains (implicit) beliefs about the following:

- Support for idea from self and others
- Habits of expressed support (self)
- Recency of visits to and from others
- Visitation habits of self (dirichlet counts)
- Affect-idea associations (dirichlet counts)

# 4. Results

Active inference allows us to formulate a normative and explainable account of cultural information spread through communication by casting cultural transmission as a bi-directional communicative process that entails a particular convergence between distinct conveyors and conveners of cultural information. We provide a proof of concept for this formalization of communication dynamics by simulating a dialogue between active inference agents holding distinct beliefs and trying to convince each other of their own beliefs.

Modeling the global dynamics of a cumulative culture (i.e., the accumulation of cultural information over a manifold of transmissions), was modeled such that—at each time point all

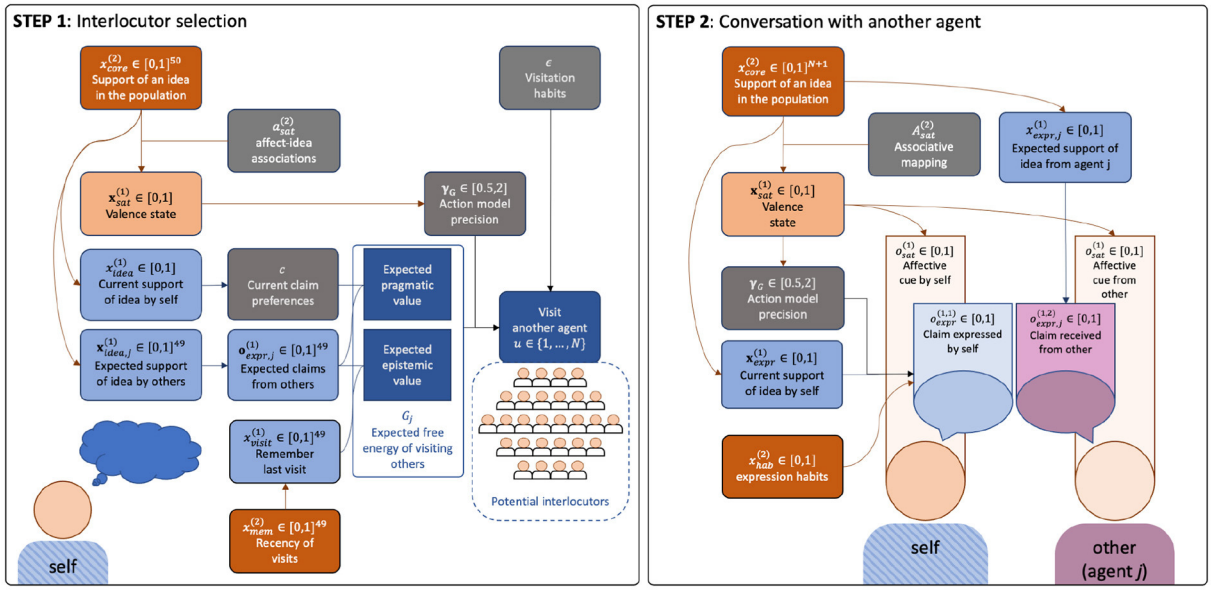



FIGURE 7
A diagram illustrating the steps of the generative model. **(Left)** Lower level Step 1: Interlocutor selection. Each day, each agent selects one interaction partner (selecting themselves means staying home). Agents cannot see each other's "opinion" before conversing. Meeting selection was conditioned on: (1) Habitual visitation drives, depending on past actions. (2) Deliberate drives, conditioned on: (2a) Expected (mis)match between expressed opinions (pragmatic value) and (2a) expected reduction in uncertainty about opinions of other agents, depending on one's memory of recent visits (epistemic value). **(Right)** Lower level Step 2: Conversation with a selected agent. Each meeting consisted in exchanges of expressed support for an idea [in the range (0,1)] and affective cues [negative-positive, in the range (0,1)]. Expressed support was conditioned on: (1) Expression habits formed during past conversations, (2) one's current support for the idea. Expressed affective cues were conditioned on one's current valence state. Affect played a role during Steps 1 and 2: Relative reliance on habitual tendencies vs. deliberation (expected free energy *G*) was regulated *via* action model precision. The latter was conditioned on one's current valence state, which was conditioned on one's current support of an idea, depending on previously learned associations between expressed ideas and concurrent affective cues (from oneself and others).

50 agents engaged in dialogue at least once (by selecting a conversation partner).

## 4.1. Local dynamics of coupled communication

### 4.1.1. The emergence of generalized synchrony from coupled communication

In nature, generalized synchrony emerges from sparse coupling between the internal states of dissipative chaotic systems (Pikovsky et al., 2003). In our model, generalized synchrony within a social system is operationalized as a convergence between belief states held by interlocutors (Figure 1). In other words, generalized synchrony between mutually inferring agents is understood as signaling a form of cultural reproduction of beliefs, namely, a mechanism by which previously distinct internal states merge and combine into one. This convergence is made possible through a particular coupling between the internal states of cultural entities, under which generalized synchrony is an emergent phenomenon. We hypothesized that without active perception and mutual model updating, belief convergence would be precluded, since interlocutors' inner states would be inaccessible to each other. That is, agents' ability to actively infer hidden states in the world and update their own model according to the sensory evidence they receive is the foundation for achieving generalized synchrony in a social system.

Our results indicate that agents' ability to listen and attune to the claims of their partner is indeed limited to the extent that they are sensitive to sensory evidence from their encultured environment (Figure 8).

To understand the implications of these findings, it is important to shed light on the way they tie in to previous work on active inference communication. In Friston and Frith (2015) provided evidence for the notion that generalized synchrony becomes altogether unattainable when agents do not possess sufficiently similar generative models. Our results go beyond this and provide evidence for the idea that only when generalized synchrony is attainable (i.e., when interlocutors possess sufficiently similar generative models), communication underlies a convergence between agents' belief states. Our simulations should therefore be understood as taking

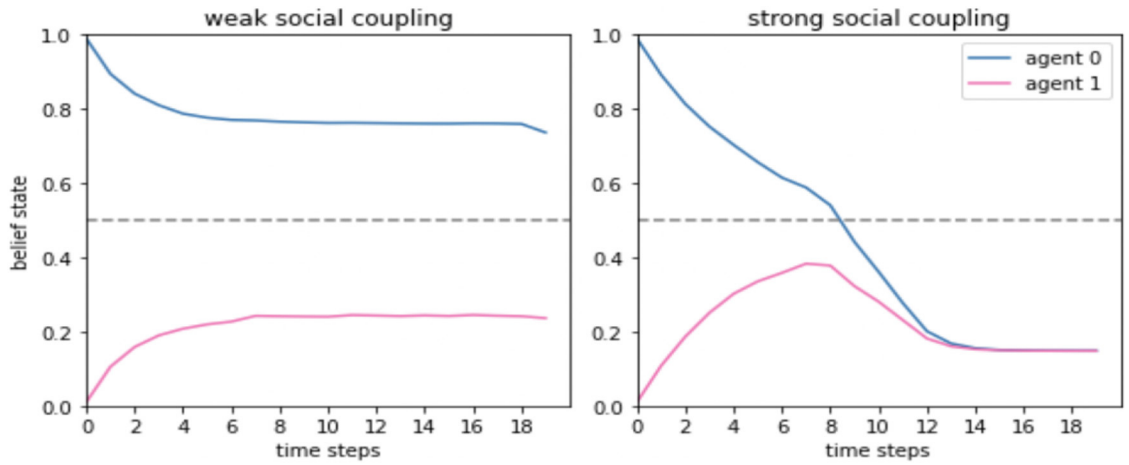



FIGURE 8
Sensitivity to observable evidence modulates the level of social coupling between agents in dialogue. This Figure and its legend were adapted from an open-source preprint of a conference paper, with permission of the authors (Figure 2 of Kastel and Hesp, 2021). In our simulations, communicated cultural information takes the form of an internal belief state held by agents with a certain probability described under the A1 matrix (Figure 3A). While this internal state is defined as a binary variable, an agent's beliefs are given by a categorical probability distribution that can take on any real number in the range (0,1). This figure shows the belief states (vertical axes) of two agents (represented in blue and pink) as they engage in dialogue across 18 time steps (horizontal axis). When the likelihood precision is low for both agents **(Left)** their internal states are very weakly coupled, such that each agent sticks to their own belief and does not attune to the claims of the other. In contrast, when both agents have high precisions **(Right)** their engagement in mutual attunement is facilitated and their beliefs converge onto one shared belief, which is then installed in both of their generative models as a shared narrative.

generalized synchrony for granted while providing evidence for the premise that the level to which agents' beliefs converge (i.e., the level of synchrony between their internal states) is modulated by their sensitivity to model evidence (*A*1).

## 4.2. Global dynamics of cumulative culture

Our simulations of a cumulative culture should be understood as modeling the dynamics of a culture that is the sum (or accumulation) of modifications to cultural beliefs and practices over time (Figure 9). While the local dyadic dynamics simulated in the previous section illustrate convergence to shared belief states held by individual agents, our global simulations leverage this synchronization to evince emergent dynamics within the population. We now review the key (predicted and) emergent phenomena we observed under this model of cumulative culture:

### 4.2.1. The introduction and spread of a novel belief induces segregation within a population

When a divergent (non status quo) belief state propagates within our synthetic population, it brings about segregation into sub-groups. Qualitatively, this is represented as a visible separation between two groups of agents: those that hold a belief that approximates the status quo (presented in red), and those that approximate the alternative, divergent belief (presented in blue).

In active inference, this communicative isolation (where agents gradually form groups of individuals they would prefer to converse with) can be explained by the attunement of interlocutor's generative models on the microscale, which translates over time—and with multiple encounters—into collective free energy minimization on the macroscale. On the microscale, local efforts to minimize free energy are expressed as agents' disinclination for meeting interlocutors that hold intractably divergent beliefs. On the macroscale, these local efforts translate into a global collective behavior of self organized separation between incongruent groups of agents, such that communicative isolation best ensures both local and collective free energy minimization. In other words, when an intractable divergent belief propagates within a homogenous population, communicative isolation between incongruent groups emerges as a strategy to minimize expected free energy, while the same strategy homogenizes the belief states of agents within congruent groups. It is interesting to reflect on the observation that the size of the two groups was roughly equivalent; a phenomena that characterizes many instances of cultural convergence (Myerson and Weber, 1993); e.g., the voting in the UK for Brexit.

### 4.2.2. Local psychological biases modulate population level segregation

The above simulations also show how differences in parameters that determine levels of confirmation bias

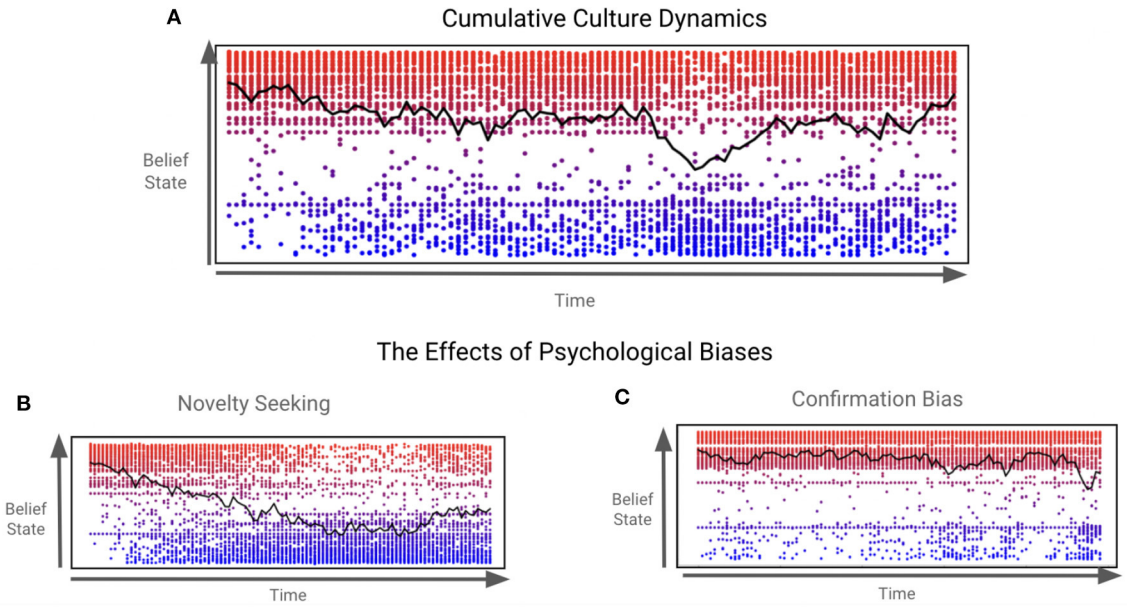



FIGURE 9
The emergence of cumulative culture. This Figure and its caption were adapted from an open-source preprint of a conference paper, with permission of the authors (Figure 3 of Kastel and Hesp, 2021). These plots depict the evolution of population-averaged support (black line) with regards to the idea that represents the initial status quo (top indicates 100% support, bottom indicates 0% support) over time (horizontal axes) along with individual core beliefs (shown in the underlying scatter plot, explained below) in three simulations, for which only the relative size of the subgroup with high confirmation bias was modulated [**(A)**: 5%, **(B)**: 15%; **(C)**: 85%]. The underlying scatter plots indicate the core beliefs of individual agents by means of their vertical location as well, with a color scale for additional clarity (red indicating maximal support for the status quo, blue maximal support for the novel idea). **(A)** Simulation of a Cumulative Culture: In this reference simulation, half of all agents are parameterized with high confirmation bias. When a divergent belief state (blue) is introduced to the status quo population (red) at the first time step, it spreads through it *via* agents in dialogue that cumulatively change the belief structure within the population. Notably, the introduction of a divergent belief seems to split the population into two subgroups: those supporting the new idea, and those adhering to the previous status quo. This effect is modulated by agents' individual strategies for choosing which interlocutors to engage with (s3). **(B)** High levels of novelty seeking in the population: When only 15% of agents are parameterized with high confirmation bias, the population exhibits high levels of novelty seeking and ends up being divided in favor of the divergent belief state, with more agents eventually holding this belief than the status quo. **(C)** High levels of confirmation bias in the population. When 85% of agents are parameterized with high confirmation bias, the population is divided in favor of the status quo belief, with more agents holding to this belief than the new and divergent belief.

(manipulated directly through $A_C^{(2)}$; see Figure 5) and novelty seeking (emerging from $G$) affect the segregation within the population into groups of agents holding either congruent beliefs or the alternative belief. When confirmation bias is relatively low (Figure 9B), the population evolves such that the majority of agents end up subscribing to the alternative belief. However, when confirmation bias is relatively high (Figure 9C), the majority of agents remains convinced of the previous status quo.

These results indicate that confirmation bias suppresses tendencies of the population as a whole toward the adoption of an idea that diverges from the status quo. When the confirmation-driven fraction of the population is relatively low (15%), we naturally observe more novelty-seeking behaviors, indicating agents are more "open-minded" and willing to meet with agents of unknown beliefs. They are intrinsically encouraged by their own curiosity to expose themselves to novel expressions. Once such agents become convinced by such interactions, they can start to popularize it for the rest of the population. If the population is, however, made up of a majority of agents driven by confirmation bias, they do not engage as much with the alternative belief and popularization is precluded.

These results are reminiscent of the widely used "adopter categories" theory, a theoretical framework outlined by Everett Rogers in his book "Diffusion of Innovations" which defines five groups in terms of their relative precession in adopting an innovation (Rogers et al., 2014). According to this framework, the first two groups to adopt an innovation are innovators and early adopters, which make up 2.5 and 13.5% of the population, respectively (Sahin, 2006). Our results appear to be consonant with the finding that relatively small numbers of early adopters and innovators play a significant role in the propagation of an innovation to other segments of the population (Dedehayir et al., 2017). One explanation for this phenomenon is that innovators and early adopters communicate innovations and their relative advantages to other segments of the population, thereby popularizing them.

## 5. Discussion

In this paper, we provide an active inference framework for the emergence of a cumulative culture from joint communication dynamics. The principal achievement of this framework is that it offers an overarching, quantitative and multiscale account against which multiple hypotheses from different domains of the social sciences may be universally tested. This accomplishment has the potentiality to bring the replication crisis faced by the psychological and social sciences in the past decade, a step closer to a resolution. A formal, standardized model of cultural evolution can evoke such an outcome as personal intuitions and culturally biased folk theories that currently make results difficult or impossible to reproduce, will become anchored to an objective and universally agreed upon verifiable account.

Notably, our framework offers a multiscale approach to the understanding of cultural evolution processes, as expressed in at least three ways. Firstly, it refers to the intrinsically hierarchical nature of the generative models themselves characterizing affective, cognitive, and behavioral dynamics (described at length in the text and summarized in Figure 2). Secondly, we combined this with the hierarchical nature of the dyadic interaction process (described at length in the text and summarized in Figure 7). Thirdly, these dyadic interactions were contextualized by a multi-agent setting for which parametrizations themselves were generated procedurally from population-level distributions of hyperparameters (described at length in the text and Appendix, illustrated in Figure 6) and including two subgroups (described in Figure 5). Our simulations depict cultural dynamics that arise from one another to form nested levels of hierarchical organization, quintessential to complex dynamical systems. This novel way of modeling cultural dynamics across layers of organization accord nicely with new approaches to artificial intelligence that originate from the notion that intelligence emerges as much from cells and societies as it does from individuals. The emerging field of biologically inspired artificial intelligence involves computational approaches that model biological systems on various layers of organization. Such artificial intelligence systems include: cellular systems; neural systems; immune systems; biomimetic, epi-genetic and evolutionary robots as well as collective systems. In this section we will discuss the specific implications of our multilevel cultural simulations on the field of biologically inspired artificial intelligence.

### 5.1. Communication models for biologically inspired artificial intelligence

Traditionally, AI has been concerned with representing the behaviors and architectures of human cognition. The preoccupation with human intelligence stems from the widely accepted notion that despite being neither the strongest nor the fastest species on earth, humans occupy a distinctly dominant position. Intellectual in nature, this dominance has previously been attributed to our culture, morality and language. However, in most of these social-cultural capacities, great apes share striking similarities with humans, yet still do not show human level intelligence, which leaves social scientists wondering about the underlying roots and causes of human intelligence. Recent studies show that despite their striking similarity to humans in most social-cultural domains, great apes are not cognitively equipped for the kinds of social coordination with others that is evident in humans (Krupenye et al., 2016; Tomasello, 2018). These findings suggest that humans might owe their remarkable intelligence to their unique ability to coordinate their behavior through joint communication and other (non verbal) cultural exchanges.

The idea that humans' cognitive skills are the result of shared intentionality, coordination, communication and social learning is known as the ontogenetic adaptation hypothesis (Tomasello, 2020). This theory stipulates that animals use social learning to gather information from their conspecifics about challenges in their environment while avoiding some of the energetic and time costs associated with a-social, trial and error learning (Clark and Dumas, 2016). According to this, social interactions- and specifically, communication and coordination- are a crucial component of human intelligence.

This makes a strong case for the use of communication models as inspiration for the development of socially intelligent artificial agents. Indeed, equipping artificial agents with the ability to accurately coordinate and communicate with other agents in their environment may well be a crucial missing piece in the modeling of advanced- human level- cognitive abilities. By modeling the underlying dynamics of social communication and coordination as we have in this paper, we bring to light an otherwise unexplored topic, which may be one of the most promising directions for achieving human level machine intelligence.

### 5.2. Cumulative culture models for biologically inspired artificial intelligence

The social sciences are on the verge of a revolution, where researchers begin to have a more complex understanding of the ways in which cultural practices and social choices interplay with, and shape human experiences. Specifically, it is becoming clear that individual intelligence is not what makes us- above other species- uniquely intelligent. Rather, the last decade has brought with it the notion that the cumulative nature of human culture is responsible for our exceptional cognitive capabilities and intelligence as a species.

This capacity to acquire- across generations- highly evolved and complex social systems such as language, cities and technologies, are said to have sharpened humans' cognitive capacities and survival strategies in a way that no other species has ever had the privilege to experience (Henrich, 2015).

Other related theories suggest that individuals are not "the brains" behind a creative idea, but that innovation is in fact a product of a collective cultural brain (Muthukrishna and Henrich, 2016). According to this, The ideas of individuals do not stand in competition or comparison with other agents in the population, but are better understood as a nexus for previously isolated ideas within it. This collective approach to cultural innovation is supported by empirical findings showing that innovation rates are higher in cultures with high sociality (i.e., large and highly interconnected populations that offer exposure to more ideas), transmission fidelity (i.e., better learning between agents) and transmission variance (i.e., a willingness to somewhat deviate from the accepted learned norms) (Muthukrishna and Henrich, 2016).

Although the capacity for cumulative culture (i.e., the capacity to acquire complex social systems through learning that accumulates across generations) in animals remains contentious (Dean et al., 2014), their expression of collective intelligent systems in swarms (Chakraborty and Kar, 2017), ant-colonies (Blum, 2005), flocks of birds (Boucherie et al., 2019), schools of fish (Boucherie et al., 2019) and other social systems, is evident in nature and has become an integral part in the field of artificial intelligence as more and more high complexity problems require bio-inspired solutions that are achievable within a reasonable period of time.

To the extent that the cumulative and collective nature of culture provides an accurate account of intelligence, as theories suggest, investigating the underlying mechanisms of intelligence may be informed by the investigation of complex social-cultural systems. In this case, providing a quantitative and measurable account of the way a "collective brain" emerges from simple, local rules of operation (namely, joint communication), as we have illustrated in this paper, becomes invaluable in the pursuit of machine intelligence.

## 5.3. Embodied active inference for biologically inspired artificial intelligence

Embodied cognition is the theory that many elements of cognition are shaped by elements of the entire body of the organism. While emphasizing the circular causality between the environment and the individual, social embodiment suggests that embodiment in social beings plays a significant role and improves upon social interactions. Justifications for social embodiment are that different body states (such as postures and facial expressions) enhance the communicative skills of embodied agents and consequently, play a central role in social information processing such that interactions between embodied agents and humans are facilitated (Bolotta and Dumas, 2021).

A natural speculation may be that robots have better skills of communication and inter-robot social inference and expression than digital avatars, since they can use their bodies for behavioral expression and coordination with other robots. However, we argue that despite lacking a physical body, active inference avatars are embodied in that the computational formalism that is applied to them (namely, active inference) implies embodiment. Technically, what this means is that active inference agents in our simulations adhere to three formal conditions for having embodied cognition:

1. They have a perceptual system which allows them to gather culturally relevant information from their surroundings. This is evident in the first layer ("perception") of the hierarchical structure of the generative model of the agents (Figure 2).
2. They have a motor system that allows them to communicate their internal states to their social environment. This is evident in the third layer ("action") of the hierarchical structure of the generative model of the agents (Figure 2).
3. They are situated in their environment such that they are able to manipulate their dynamic surroundings through their actions. This is evident in agents' ability to listen and attune to each others' belief expression in a way that allows for a coupling of their internal states and the emergence of generalized synchrony between them (Figure 8).

The fact that agents under the active inference formulation conform to these three conditions is non trivial, and it points to the fact that these agents could not be simply replaced by any hypothetical- non embodied- simulated intelligent being. In other words, our simulations would not make sense unless applied to a population that adheres to the certain criteria aforementioned. We could only apply our simulations to agents that adhere to all these conditions (i.e., have embodied cognition), or our simulations simply would not work. In this sense, our agents may not be physically embodied robots, but we argue that- by definition- as active inference agents capable of perceptions and actions in a situated environment, they are software embodied agents. Had we put this software into social robots that had the hardware equivalent of "ears" and "mouths," we would be able to produce embodied robots in a way that would improve their social interactions. Crucially, we argue here that embodiment must be present in both the software and the hardware for social interactions of agents to be enhanced by it, and that the active inference formalism implies embodiment for the former.

## 5.4. Limitations and future research

Although this paper provides important insights into the underlying dynamics of social-cultural systems, it entails certain limitations that will now be outlined and may be addressed in future research.

First, our communication simulations assume that social exchange is limited to a dyad, when in fact generalized synchrony in nature may occur between multiple coupled systems. Our formulations of the exchange of social information as communication therefore represent only a specific case of generalized synchrony that might highlight a much more encompassing phenomenon. As an example, through social media, cultural information can reach large populations at a given time point. The idea that generalized synchrony between inferring agents may go beyond the emergent behavior of two communicators and exist between ensembles of coupled self organizing systems has also been considered in the active inference literature (Palacios et al., 2019).

Second, while we provide a formulation of the way modifications to cultural information occur during communication (i.e., the transmission of social information) and we have simulated the emergence of cumulative culture from these dynamics (i.e., the prevalence of social information), we have not provided an account of the way novel social information is introduced into a population to begin with. We have assumed that belief states are gradually modified with every cultural exchange, such that the outcome of this exchange may be considered novel by virtue of it being a unique recombination of existing beliefs and practices. Future research may focus on asking important questions like: Why are we inclined to say that innovation is a unique event that does not occur with every cultural transmission? More importantly, how can we define and even model the difference between a slight modification to a cultural trait and innovation?

The importance of identifying exactly what constitutes innovation and how to model its emergence is critical for an accurate understanding of socio-cultural dynamics because it would bring the circular dynamics of a complex culture to a required close (Figure 10). Under such an account, not only would cumulative culture naturally emerge from a complex network of agents engaged in joint communication (as shown in this proposal), but innovation would emerge from cumulative culture and underlie communication in a repeating, recursive loop that is the hallmark of complex dynamical systems.

In the simulation environment presented here, efficient communication could be considered, to some extent, as reflecting local communicative needs and these needs are grounded in properties of this socio-cultural environment. However, the environment itself does not impose any of the practical constraints known to drive real-life human behaviors (e.g., need for food, warmth, hygiene). Previously, researchers have argued that practical benefits in adapting to the environment tend to accelerate the repetition and widespread adoption of cultural practices (Kashima et al., 2019). In future work, the authors aim to expand on these notions by enriching the simulated environment with actual practical constraints.

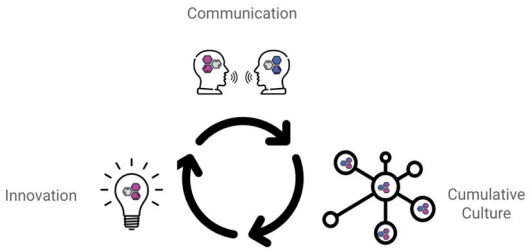



FIGURE 10
The circular dynamics of cultural evolution.

In the current work, emotional valence was tied to action confidence in a top-down manner (it affected action-model precision for both meeting selection and expression) but not in a bottom-up manner (e.g., based on action outcomes). For the sake of simplicity, influences on emotional valence were purely associative (i.e., based on emotional expressions of conversation partners). Therefore, one natural and valuable extension of these simulations in future work would be to fully incorporate the recursive and principled formulation of emotional valence that has been derived from deep active inference (Hesp et al., 2021), which naturally tracks changes in subjective fitness. This recurrent formulation, in particular when combined with imagination-induced affect (see Hesp et al., 2020), will specifically benefit from the grounding of these simulated cultural exchanges in a more elaborate virtual environment combined with agents that have actual bodily and social needs such that subjective fitness estimates (based on action-model precision) come to confer some practical relevance (as described in the preceding paragraph).

Finally, another limitation of our simulations is that the agents' freedom for choosing the interlocutors they want to engage with, might bias cultural transmission in a way that does not apply to some forms of social interaction. Specifically, meeting selection, or the freedom to voluntarily select the transmitting interlocutor, does not extend to social interactions in which agents do not have a choice in determining the source of their cultural learning. For example, during development, children are constantly exposed to individuals, social situations, cultural practices, and conversations that they do not voluntarily select. In this case, it is the Parents' culturally dominant behaviors that play a central role in the development of children's internalization of cultural beliefs, rather than the voluntary actions of their children (Fernald and Morikawa, 1993; Senzaki et al., 2016).

## 6. Conclusion

In this paper, we employed a Bayesian framework—known as active inference—to formally account for the dynamics underlying (local) communication and (global) cumulative culture dynamics, thus contributing to the ever-growing body of research on multi-agent Bayesian models (e.g., Gunji et al., 2018) and collective active inference (e.g., Friedman et al., 2021; Heins et al., 2022) Under our account, the social "transmission" of cultural information has been cast as a fundamentally bidirectional process of communication, which has been shown in the previous active inference literature to induce a generalized synchrony between the internal (belief) states of agents holding sufficiently similar generative models. Building on this work, we operationalized generalized synchrony as a particular convergence between the internal states of interlocutors, and show that it depends sensitively on the precision of observation or likelihood mappings in a generative model of communicative exchange. When we simulate a population of agents that simultaneously engage in communication over time, cumulative culture emerges as the collective behavior brought about by local belief updating (active inference and learning in a dyadic setting). Our simulations show that when a divergent belief is introduced to the status quo, it spreads within the population and brings about a collective behavior characterized by a certain degree of segregation between different belief groups. The level to which the status quo population defects to the divergent belief is mediated by local psychological biases for confirmation bias (as directly manipulated) and novelty seeking (as emergent from procedural generation of parameters). These cultural (c.f., voting) equilibria are minimizers of collective or joint free energy that emerge from the imperative to minimize uncertainty and surprise in dyadic exchanges.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

NK and CH implemented the formalism of active inference, conducted the simulations, and made the figures. NK developed the theoretical background for the manuscript and wrote up the first draft of manuscript. KR and CH edited the manuscript and linked components in the literature review with prior work in the field. CH and KF edited the manuscript and further developed the formalism of active inference in the methods, generative model, and results sections. All authors contributed to the article and approved the submitted version.

## Funding

This research was undertaken thanks in part to funding from a NWO Research Talent Grant of the Dutch Government (CH; No. 406.18.535) and by a Wellcome Trust Principal Research Fellowship (KF; Ref. 088130/Z/09/Z).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2022.944986/full#supplementary-material

## References

Bettencourt, L. M., Cintrón-Arias, A., Kaiser, D. I., and Castillo-Chávez, C. (2006). The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. *Physica A* 364, 513–536. doi: 10.1016/j.physa.2005.08.083

Blum, C. (2005). Ant colony optimization: introduction and recent trends. *Phys. Life Rev*. 2, 353–373. doi: 10.1016/j.plrev.2005.10.001

Bolotta, S., and Dumas, G. (2021). Social neuro AI: social interaction as the" dark matter" of AI. *arXiv preprint arXiv:2112.15459*. doi: 10.3389/fcomp.2022.846440

Boucherie, P. H., Loretto, M. C., Massen, J. J., and Bugnyar, T. (2019). What constitutes "social complexity" and "social intelligence" in birds? Lessons from ravens. *Behav. Ecol. Sociobiol*. 73, 1–14. doi: 10.1007/s00265-018-2607-2

Buskell, A., Enquist, M., and Jansson, F. (2019). A systems approach to cultural evolution. *Palgrave Commun*. 5, 1–15. doi: 10.1057/s41599-019-0343-5

Chakraborty, A., and Kar, A. K. (2017). "Swarm intelligence: a review of algorithms," in *Nature-Inspired Computing and Optimization*, 475–494.

Clark, H. H., Brennan, S. E. (1991). "Grounding in communication," in *Perspectives on Socially Shared Cognition*, eds L. B. Resnick, J. M. Levine, and S. D. Teasley (American Psychological Association), 127–149.

Clark, I., and Dumas, G. (2016). The regulation of task performance: a trans-disciplinary review. *Frontiers in psychology* 6, 1862. doi: 10.3389/fpsyg.2015.01862

Constant, A., Ramstead, M. J., Veissiere, S. P., Campbell, J. O., and Friston, K. J. (2018). A variational approach to niche construction. *J. R. Soc. Interface* 15, 20170685. doi: 10.1098/rsif.2017.0685

Constant, A., Ramstead, M. J., Veissière, S. P., and Friston, K. (2019). Regimes of expectations: an active inference model of social conformity and human decision making. *Front. Psychol.* 10, 679. doi: 10.3389/fpsyg.2019.00679

Creanza, N., Kolodny, O., and Feldman, M. W. (2017). Cultural evolutionary theory: how culture evolves and why it matters. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7782–7789. doi: 10.1073/pnas.1620732114

Dean, L. G., Vale, G. L., Laland, K. N., Flynn, E., and Kendal, R. L. (2014). Human cumulative culture: a comparative perspective. *Biol. Rev.* 89, 284–301. doi: 10.1111/brv.12053

Dedehayir, O., Ortt, R. J., Riverola, C., and Miralles, F. (2017). Innovators and early adopters in the diffusion of innovations: a literature review. *Int. J. Innovat. Manag.* 21, 1740010. doi: 10.1142/S1363919617400102

Derex, M., and Mesoudi, A. (2020). Cumulative cultural evolution within evolving population structures. *Trends Cogn. Sci.* 24, 654–667. doi: 10.1016/j.tics.2020.04.005

Dunstone, J., and Caldwell, C. A. (2018). Cumulative culture and explicit metacognition: a review of theories, evidence and key predictions. *Palgrave Commun.* 4, 1–11. doi: 10.1057/s41599-018-0200-y

Echterhoff, G., Higgins, E. T., and Levine, J. M. (2009). Shared reality: experiencing commonality with others' inner states about the world. *Perspect. Psychol. Sci.* 4, 496–521. doi: 10.1111/j.1745-6924.2009.01161.x

Fernald, A., and Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Dev.* 64, 637–656. doi: 10.2307/1131208

Friedman, D. A., Tschantz, A., Ramstead, M. J. D., Friston, K., and Constant, A. (2021). Active Inferants: an active inference framework for ant colony behavior. *Front. Behav. Neurosci.* 15, 647732. doi: 10.3389/fnbeh.2021.647732

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., and Frith, C. (2015). A duet for one. *Conscious Cogn.* 36, 390–405. doi: 10.1016/j.concog.2014.12.003

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Gunji, Y.-P., Murakami, H., Tomaru, T., and Basios, V. (2018). Inverse Bayesian inference in swarming behaviour of soldier crabs. *Phil. Trans. R. Soc. A* 376, 20170370. doi: 10.1098/rsta.2017.0370

Haugland, S. W. (2021). The changing notion of chimera states, a critical review. *J. Phys. Complex.* 2 032001doi: 10.1088/2632-072X/ac0810

Heins, C., Klein, B., Demekas, D., Aguilera, M., and Buckley, C. (2022). Spin glass systems as collective active inference. *arXiv preprint arXiv:2207.06970*. doi: 10.48550/arXiv.2207.06970

Henrich, J. (2015). "The secret of our success," in *The Secret of Our Success* (Princeton, NJ: Princeton University Press).

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., and Friston, K. (2019). "A multi-scale view of the emergent complexity of life: a free-energy proposal," in *Evolution, Development and Complexity* (Cham: Springer), 195–227.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. D. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Comput.* 33, 398–446. doi: 10.1162/neco_a_01341

Hesp, C., Tschantz, A., Millidge, B., Ramstead, M., Friston, K., and Smith, R. (2020). "Sophisticated affective inference: Simulating anticipatory affective dynamics of imagining future events," in *Communications in Computer and Information Science, Vol. 1326*, eds T. Verbelen, P. Lanillos, and C. Buckley (Cham: Springer Nature), 179–186.

Jansson, F., Aguilar, E., Acerbi, A., and Enquist, M. (2021). Modelling cultural systems and selective filters. *Philos. Trans. Royal Soc. B. Biol. Sci.* 376, 20200045. doi: 10.1098/rstb.2020.0045

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.* 112, 323–343. doi: 10.1007/s00422-018-0753-2

Kashima, Y., Bain, P. G., and Perfors, A. (2019). The psychology of cultural dynamics: what is it, what do we know, and what is yet to be known? *Annu Rev. Psychol.* 70, 499–529. doi: 10.1146/annurev-psych-010418-103112

Kashima, Y., Kirley, M., Stivala, A., and Robins, G. (2017). "Modeling cultural dynamics," in *Computational Social Psychology* (Routledge), 281–307.

Kashima, Y., Perfors, A., Ferdinand, V., and Pattenden, E. (2021). Ideology, communication and polarization. *Philos. Trans. R. Soc. B* 376, 20200133. doi: 10.1098/rstb.2020.0133

Kastel, N., and Hesp, C. (2021). "Ideas worth spreading: A free energy proposal for cumulative cultural dynamics," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Cham: Springer), 784–798.

Kempe, M., and Mesoudi, A. (2014). Experimental and theoretical models of human cultural evolution. *Wiley Interdiscip. Rev. Cogn. Sci.* 5, 317–326. doi: 10.1002/wcs.1288

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science* 354, 110–114. doi: 10.1126/science.aaf8110

Migliano, A. B., and Vinicius, L. (2022). The origins of human cumulative culture: from the foraging niche to collective intelligence. *Philos. Trans. Royal Soc. B. Biol. Sci.* 377, 20200317. doi: 10.1098/rstb.2020.0317

Muthukrishna, M., and Henrich, J. (2016). Innovation in the collective brain. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20150192. doi: 10.1098/rstb.2015.0192

Myerson, R. B., and Weber, R. J. (1993). A theory of voting equilibria. *Am. Polit. Sci. Rev.* 87, 102–114. doi: 10.2307/2938959

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220.

Nicol, C. J. (1995). The social transmission of information and behavior. *Appl. Anim. Behav. Sci.* 44, 79–98. doi: 10.1016/0168-1591(95)00607-T

Nicolis, G., and Nicolis, C. (2012). *Foundations of Complex Systems: Nonlinear Dynamics, Statistical Physics, Information and Prediction*," *World Scientific, 2nd Edn*.

Palacios, E. R., Isomura, T., Parr, T., and Friston, K. (2019). The emergence of synchrony in networks of mutually inferring neurons. *Scient. Rep.* 9, 1–14. doi: 10.1038/s41598-019-42821-7

Pikovsky, A., Kurths, J., Rosenblum, M., and Kurths, J. (2003). *Synchronization: A Universal Concept in Nonlinear Sciences (No. 12)*. Cambridge: Cambridge University Press.

Rogers, E. M., Singhal, A., and Quinlan, M. M. (2014). "Diffusion of innovations," in *An integrated Approach to Communication Theory and Research* (London: Routledge), 432–448.

Russell, J. A., and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *J. Pers. Soc. Psychol.* 76, 805–819. doi: 10.1037//0022-3514.76.5.805

Sahin, I. (2006). Detailed review of Rogers' diffusion of innovations theory and educational technology-related studies based on Rogers' theory. *Turkish Online J. Educ. Technol.* 5, 14–23.

Senzaki, S., Masuda, T., Takada, A., and Okada, H. (2016). The communication of culturally dominant modes of attention from parents to children: a comparison of Canadian and Japanese parent-child conversations during a joint scene description task. *PLoS ONE* 11, e0147199. doi: 10.1371/journal.pone.0147199

Smith, R., Parr, T., and Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Front. Psychol.* 10, 2844. doi: 10.3389/fpsyg.2019.02844

Stout, D., and Hecht, E. E. (2017). Evolutionary neuroscience of cumulative culture. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7861–7868. doi: 10.1073/pnas.1620738114

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proc. Natl. Acad. Sci. U.S.A.* 115, 8491–8498. doi: 10.1073/pnas.1804761115

Tomasello, M. (2020). "Why don't apes point?," in *Roots of Human Sociality* (London: Routledge), 506–524.

Veissière, S., Constant, A., Ramstead, M., Friston, K., and Kirmayer, L. (2020). Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 43, E90. doi: 10.1017/S0140525X19001213

Zakharova, A. (2020). *Chimera Patterns in Networks: Interplay between Dynamics, Structure, Noise, and Delay*. Cham: Springer Nature.

# Frontiers in Neurorobotics

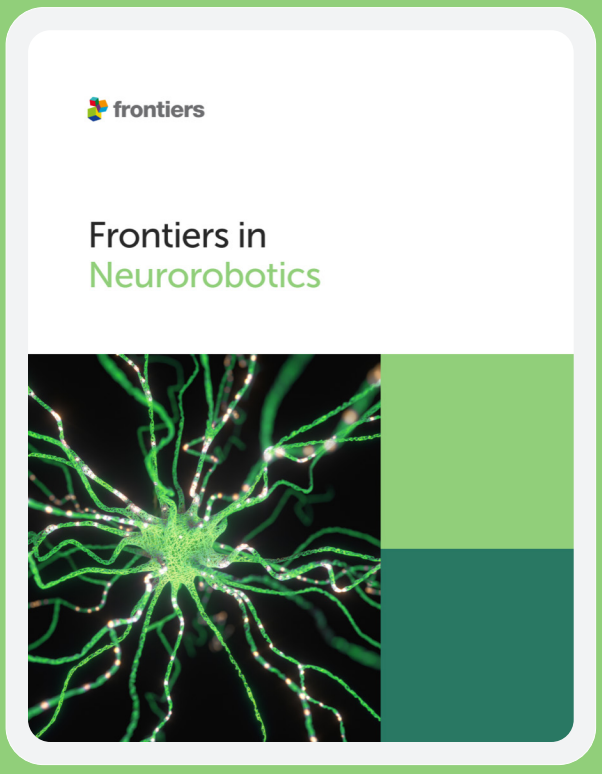
**Investigates embodied autonomous neural systems and their impact on our lives**

Part of the most cited neuroscience series, this journal advances understanding of neurorobotics - from prosthetic devices to brain machine interfaces, and wearable systems to home appliances.

---

## Discover the latest Research Topics

See more →




**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact