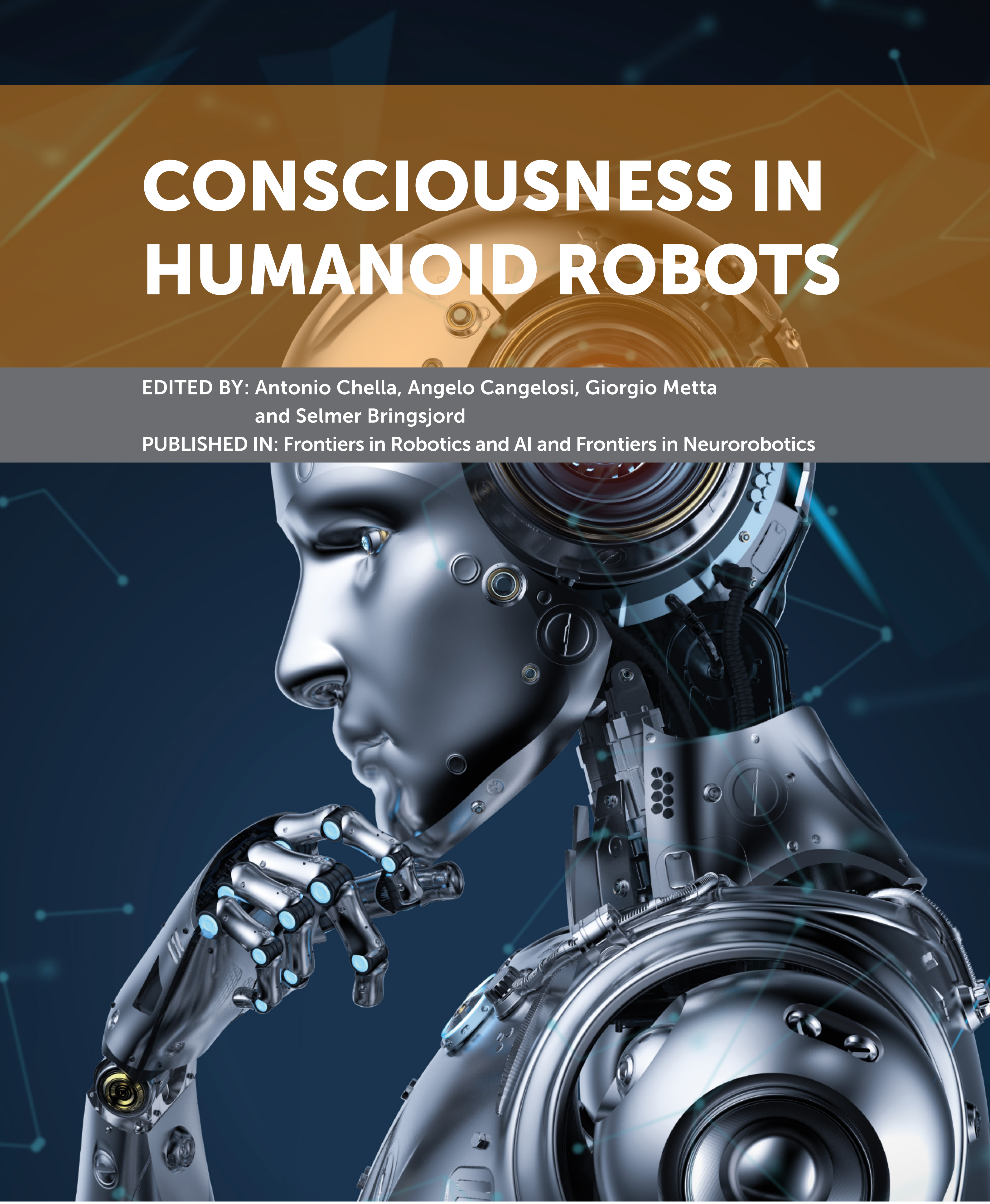
CONSCIOUSNESS IN HUMANOID ROBOTS

EDITED BY: Antonio Chella, Angelo Cangelosi, Giorgio Metta
and Selmer Bringsjord

PUBLISHED IN: Frontiers in Robotics and AI and Frontiers in Neurorobotics



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-866-0
DOI 10.3389/978-2-88945-866-0

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: researchtopics@frontiersin.org

CONSCIOUSNESS IN HUMANOID ROBOTS

Topic Editors:
Antonio Chella, Università degli Studi di Palermo, Italy; ICAR-CNR, Italy
Angelo Cangelosi, University of Manchester, United Kingdom
Giorgio Metta, Istituto Italiano di Tecnologia, Italy
Selmer Bringsjord, Rensselaer Polytechnic Institute, United States

Cover image: Phonlamai Photo/Shutterstock.com

Building a conscious robot is a scientific and technological challenge. Debates about the possibility of conscious robots and the related positive outcomes and hazards for human beings are today no longer confined to philosophical circles.

Robot consciousness is a research field aimed at a two-part goal: on the one hand, researchers working in robot consciousness take inspiration from biological consciousness to build robots that present forms of experiential and functional consciousness. On the other hand, scholars employ robots as tools to better understand biological consciousness.

Thus, part one of the goal concerns the replication of aspects of biological consciousness in robots, by unifying a variety of approaches from AI and robotics, cognitive robotics, epigenetic and affective robotics, situated and embodied robotics, developmental robotics, anticipatory systems, and biomimetic robotics.

Part two of the goal is pursued by employing robots to advance and mark progress in the study of consciousness in humans and animals. Notably, neuroscientists involved in the study of consciousness do not exclude the possibility that robots may be conscious.

This eBook comprises a collection of thirteen manuscripts and an Editorial published by Frontiers in Robotics and Artificial Intelligence, under the section Humanoid Robotics, and Frontiers in Neurorobotics, on the topic "Consciousness in Humanoid Robots." This compendium aims at collating the most recent theoretical studies, models, and case studies of machine consciousness that take the humanoid robot as a frame of reference. The content in the articles may be applied to many different kinds of robots, and to software agents as well.

Citation: Chella, A., Cangelosi, A., Metta, G., Bringsjord, S., eds. (2019). Consciousness in Humanoid Robots. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-866-0

Table of Contents

EDITORIAL
published: 22 March 2019
doi: 10.3389/frobt.2019.00017

Editorial: Consciousness in Humanoid Robots

Antonio Chella[1,2]*, *Angelo Cangelosi*[3], *Giorgio Metta*[4] and *Selmer Bringsjord*[5,6]

[1] RoboticsLab, Department of Industrial and Digital Innovation, University of Palermo, Palermo, Italy, [2] Cognitive Robotics and Social Sensing Laboratory, ICAR-CNR, Palermo, Italy, [3] School of Computer Science, The University of Manchester, Manchester, United Kingdom, [4] iCub Facility, Istituto Italiano di Tecnologia, Genova, Italy, [5] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, United States, [6] Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY, United States

Keywords: robot consciousness, robot awareness, machine consciousness, artificial consciousness, autonomic robotics

Editorial on the Research Topic

Consciousness in Humanoid Robots

Building a conscious robot is a grand scientific and technological challenge. Debates about the possibility of conscious robots and the related positive outcomes and hazards for human beings are today no longer confined to philosophical circles.

There is no accepted definition of consciousness: see Vimal (2009) for an overview of different meanings of the word. However, it is useful to point out the distinction of consciousness as experience and consciousness as function. From the point of view of experience, a subject is conscious when she feels visual experiences, bodily sensations, mental images, emotions (Chalmers, 1995). As Nagel (1974) points out, a subject has conscious experience if there is something it is like to be that subject. From the point of view of function, a conscious subject is able to process information which is globally available (Dehaene et al., 2017), she integrates information (Tononi, 2008), she is introspectively aware of herself (Floridi, 2005). Moreover, she generates inner speech (Morin, 2005), she possesses an inner model of herself and external environment (Holland, 2003), she is able to anticipate perceptual and behavioral activities (Hesslow, 2002), and she acts by sensorimotor interactions with the external world (O'Regan and Noë, 2001).

Bringsjord (2007) contrasts the possibility of experiences in robots and proposes the notion of cognitive consciousness (Bringsjord et al., 2018), offering a definition in terms of formal axioms. Bringsjord et al. (2015) report the best example of cognitive consciousness by discussing a robot that passed the human test of self-consciousness proposed by Floridi (2005).

Robot consciousness is a research field aimed at two-fold goal: on the one side, scholars working in robot consciousness take inspiration from biological consciousness to build robots that present forms of experiential and functional consciousness. On the other side, scholars employ robots as tools to better understand biological consciousness.

Thus, a goal concerns the replication of aspects of biological consciousness in robots, by unifying a variety of approaches from AI and robotics, cognitive robotics, epigenetic and affective robotics, situated and embodied robotics, developmental robotics, anticipatory systems, and biomimetic robotics (Chella and Manzotti, 2009; Bringsjord and Govindarajulu, 2018).

The other goal of robot consciousness concerns the employment of robots to mark progress in the study of consciousness in humans and animals. Notably, neuroscientists involved in the study of consciousness do not exclude the possibility that robots may be conscious (Dehaene et al., 2017).

OPEN ACCESS

Edited by:
Nadia Magnenat Thalmann,
Université de Genève, Switzerland

Reviewed by:
Rodolphe Gelin,
Softbank Robotics, France

***Correspondence:**
Antonio Chella
antonio.chella@unipa.it

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *31 October 2018*
Accepted: *28 February 2019*
Published: *22 March 2019*

Citation:
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

This e-book comprises a collection of 13 manuscripts published by Frontiers in Robotics and Artificial Intelligence, under the section Humanoid Robotics, on the topic on "Consciousness in Humanoid Robots." This compendium aims at collating the most recent theoretical studies, models, and case studies of machine consciousness that take the humanoid robot as a frame of reference. However, the arguments of the articles may be applied to different kinds of robots and even to software agents.

OVERVIEW OF THE CONTENTS OF THE E-BOOK

A methodological strategy for the study of robot consciousness is introduced by Reggia et al. by means of the concept of a computational correlate of consciousness. This parallels the concept of a neural correlate of consciousness in the brain. Thus, they describe a cognitive robot able to learn by imitation through low-level cognitive components such as working memory and causal reasoning mechanisms. The top-down cognitive control of the working memory of the robot is a potential computational correlate of robot consciousness.

According to Manzotti and Chella, the typical approaches toward robot consciousness as, for example, global workspace, information integration, enaction, cognitive mechanisms, embodiment, constitute the Good Old-Fashioned Artificial Consciousness. These share the same conceptual fallacy that the authors name "the intermediate level fallacy." Thus, they outline a new conceptual framework toward robot consciousness.

The attentional mechanisms, theory of mind, and the role of emotions are all critical aspects in the study of the mechanisms underlying consciousness in humans and in robots. In this context, Graziano proposes a theory based on the attention schema as a starting point to build a conscious robot. The attention schema theory may explain how an entity lays claim to possess subjective awareness. According to Graziano, it is possible to create a robot with a rich internal model of consciousness that attributes consciousness to itself and to the people it interacts with, and that uses this attribution to predict human behavior.

Winfield proposes an artificial theory of mind that would provide robots with new capabilities related to social intelligence for human-robot interaction. The author suggests that a simulation-based internal model may offer a new basis for the artificial theory of mind. Internal models equip the robot with a model of itself and the environment, including other agents, so that the robot can test its possible actions and anticipate the consequences for itself and the other agents.

Cominelli et al. present the cognitive system SEAI (Social Emotional Artificial Intelligence) aimed for social and emotional robots designed as a bio-inspired system with a model of emotion and reasoning capabilities. In particular, SEAI comprises a simulation of Damasio's theory of consciousness.

Wang et al. and Chatila et al. consider the relevant problem of robot self-consciousness. In details, Wang et al. discuss self-consciousness in terms of NARS, an implemented general-purpose intelligent system. The authors explain how a general-intelligent system needs a notion of the "self" based on the experiences accumulated by the system during its development. The implementation of self-awareness and self-control capabilities in NARS is at an early stage; however, the overall design fits well with the processes in the human mind.

According to Chatila et al., the self-consciousness of a robot emerges by the distinction operated by the robot between its own body and the external environment. The paper proposes a cognitive architecture that considers several aspects: the perception of the robot; the interaction capabilities with the external environment; the learning phase; the interaction with other agents; the decision-making capacities.

Aspects related to architectural features for a conscious robot have been treated by Kinouchi and Mackin, Van de Velde, and Balkenius et al. In particular, Kinouchi and Mackin propose a cognitive neural architecture for a conscious robot where the primary role of consciousness is the adaptation at the system-level. The proposed architecture is based on a two-level design: the first level is related to awareness, habitual behavior, and the binding problem. The second level is associated with the general goal-directed behavior of the robot.

Van de Velde provides suggestions for robot architectures by analyzing the roles of cognitive processing and access consciousness in the brain. The author argues that consciousness is a process which is referred to *in situ* representations in the brain that underlie the possibility of cognitive access. Given this, consciousness may be related to a continuous process of cognitive access controlled by the activity of *in situ* representations themselves, as in the operations of queries and answers.

Balkenius et al. discuss the roles of memory and the inner world for a conscious robot. The authors introduce a memory model, based on neurophysiological data, that considers many aspects, such as object permanence and episodic memory. The three components of the model are an identification network, a localization network, and a working memory network. The mechanisms that fill in the sensations to the generation of perceptions can be detached from sensory input and run in isolation, allowing for planning mechanisms and daydreaming.

The active inference framework is discussed in detail by Linson et al. and by Biehl et al. The active inference framework is a bridge between computational neuroscience and robotics to psychology and phenomenology. The framework provides a theoretical basis for a unified treatment of particles, organisms, and interactive machines. The theory considers perception, reasoning, and action selection under the heading of a single principle. Notably, it suggests biologically plausible explanations for cognitive phenomena and implications for robot consciousness.

Finally, Signorelli analyses some misconceptions related to the next generations of conscious robots. The author discusses the sense in which a robot could reach capabilities at the human level, asserting that it could be possible only in case of a sentient robot. Then, a robot would be classified according to the human types of cognition. An important aspect of the author's discussion is that a conscious robot would not overcome humans but, on the contrary, it could present the very same limitations presented by humans.

CONCLUSIONS

In summary, the advent of a conscious robot would be a tremendous scientific and technological leap.

The 13 contributions collected in this e-book touch essential aspects of the current debate about robot consciousness as the relationship between phenomenology and cognition, the role of theory of mind and self-awareness, the roles of attention and emotions, the possible problems arising from a conscious robot among us. Insights concerning the design of cognitive architectures and initial implementations are discussed. The active inference framework is investigated as a promising general theory able to consider biological and robot consciousness.

The main message from this e-book is the need for tight relationships between scientific and technological research on robot consciousness and understanding of the processes related to biological consciousness. In fact, understanding the underlying aspects of biological consciousness would greatly help to build a new generation of conscious robots, which, in turn, would contribute to a better understanding of biological consciousness.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Bringsjord, S. (2007). Offer: one billion dollars for a conscious robot. If you're honest, you must decline. *J. Conscious. Stud.* 14, 28–43.

Bringsjord, S., Bello, P., and Govindarajulu, N. S. (2018). "Toward axiomatizing consciousness," in *The Bloomsbury Companion to the Philosophy of Consciousness,* ed D. Jacquette (London: Bloomsbury Academic), 289–324.

Bringsjord, S., and Govindarajulu, N. S. (2018). "Artificial intelligence," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta. Available online at: https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R., and Sen, A. (2015). "Real robots that pass tests of self-consciousness," in *Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)* (New York, NY: IEEE), 498–504. doi: 10.1109/ROMAN.2015.7333698

Chalmers, D. (1995). Facing up to the problem of consciousness. *J. Conscious. Stud.* 2, 200–219.

Chella, A., and Manzotti, R., (2009). Machine consciousness: a manifesto for robotics. *Int. J. Mach. Conscious.* 1, 33–51. doi: 10.1142/S1793843009000062

Dehaene, S., Lau, H., Kouider, S. (2017). What is consciousness, and could machines have it? *Science* 358, 486–492. doi: 10.1126/science.aan8871

Floridi, L. (2005). Consciousness, agents and the knowledge game. *Mind Mach.* 15, 415–444. doi: 10.1007/s11023-005-9005-z

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends Cogn. Sci.* 6, 242–247. doi: 10.1016/S1364-6613(02)01913-7

Holland, O. (2003). Robots with internal models – a route to machine consciousness? *J. Conscious. Stud.* 10, 77–109.

Morin, A. (2005). Possible links between self-awareness and inner speech. *J. Conscious. Stud.* 12, 115–134.

Nagel, T. (1974). What is like to be a bat? *Philos. Rev.* 83, 435–450. doi: 10.2307/2183914

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

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242. doi: 10.2307/25470707

Vimal, R. L. P. (2009). Meaning attributed to the term 'consciousness' – an overview. *J. Conscious. Stud.* 16, 9–27.

Conflict of Interest Statement: 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.

HYPOTHESIS AND THEORY
published: 14 November 2017
doi: 10.3389/frobt.2017.00060

The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness

*Michael S. A. Graziano**

Department of Psychology and Neuroscience, Princeton University, Princeton, NJ, United States

The purpose of the attention schema theory is to explain how an information-processing device, the brain, arrives at the claim that it possesses a non-physical, subjective awareness and assigns a high degree of certainty to that extraordinary claim. The theory does not address how the brain might actually *possess* a non-physical essence. It is not a theory that deals in the non-physical. It is about the computations that cause a machine to make a claim and to assign a high degree of certainty to the claim. The theory is offered as a possible starting point for building artificial consciousness. Given current technology, it should be possible to build a machine that contains a rich internal model of what consciousness is, attributes that property of consciousness to itself and to the people it interacts with, and uses that attribution to make predictions about human behavior. Such a machine would "believe" it is conscious and act like it is conscious, in the same sense that the human machine believes and acts.

Keywords: attention, awareness, body schema, internal model, visual attention

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo,
Italy

Reviewed by:
Ronald A. Rensink,
University of British Columbia,
Canada
Paul Syverson,
United States Naval Research
Laboratory, United States

****Correspondence:***
Michael S. A. Graziano
graziano@princeton.edu

Specialty section:
This article was submitted to
Humanoid Robotics,
a section of the journal
Frontiers in Robotics and AI

Received: *16 August 2017*
Accepted: *01 November 2017*
Published: *14 November 2017*

Citation:
Graziano MSA (2017) The Attention
Schema Theory: A Foundation for
Engineering Artificial Consciousness.
Front. Robot. AI 4:60.
doi: 10.3389/frobt.2017.00060

INTRODUCTION

This article is part of a special issue on consciousness in humanoid robots. The purpose of this article is to summarize the attention schema theory (AST) of consciousness for those in the engineering or artificial intelligence community who may not have encountered previous papers on the topic, which tended to be in psychology and neuroscience journals. The central claim of this article is that AST is mechanistic, demystifies consciousness and can potentially provide a foundation on which artificial consciousness could be engineered. The theory has been summarized in detail in other articles (e.g., Graziano and Kastner, 2011; Webb and Graziano, 2015) and has been described in depth in a book (Graziano, 2013). The goal here is to briefly introduce the theory to a potentially new audience and to emphasize its possible use for engineering artificial consciousness.

The AST was developed beginning in 2010, drawing on basic research in neuroscience, psychology, and especially on how the brain constructs models of the self (Graziano, 2010, 2013; Graziano and Kastner, 2011; Webb and Graziano, 2015). The main goal of this theory is to explain how the brain, a biological information processor, arrives at the claim that it possesses a non-physical, subjective awareness and assigns a high degree of certainty to that extraordinary claim. The theory does not address how the brain might actually *possess* a non-physical essence. It is not a theory that deals in the non-physical. It is about the computations that cause a machine to make a claim and to assign a high degree of certainty to the claim. The theory is in the realm of science and engineering.

Given a mechanistic theory of this type, my best guess is that artificial consciousness will arrive relatively soon, within the next century, and that even farther down the road people will be able to migrate their minds to new hardware much like we now migrate essential data and algorithms from an obsolete computer to an upgraded model. That type of technology will obviously be transformational, though whether good or bad I am not sure. Every aspect of human existence—culture, politics, health, preservation of knowledge and wisdom across periods of time, human dispersion across space, and other environments hostile to biology—will be fundamentally changed by the easy transferability of minds to new hardware. As crazily science fiction as these possibilities sound, I see our technology moving in that direction. My hope is that AST will provide some initial insights into consciousness that are concrete enough, and mechanistic enough, that engineers can build upon it to facilitate the technology.

THE CRUCIAL DIFFERENCE BETWEEN MIND AND LAPTOP

Before explaining the theory, it is useful to specify what phenomenon it purports to tackle. The term consciousness, after all, has many, sometimes conflicting meanings. To help specify the meaning used here, consider the difference between a brain and a modern personal computer. Of course there are many differences, but one seems more consequential than others. The brain has a subjective experience associated with a subset of the information that it processes.

You can connect a computer to a camera and program it to process visual information—color, shape, size, and so on. The human brain does the same, but in addition, we report a subjective experience of those visual properties. This subjective experience is not always present. A great deal of visual information enters the eyes, is processed by the brain and even influences our behavior through priming effects, without ever arriving in awareness. Flash something green in the corner of vision and ask people to name the first color that comes to mind, and they may be more likely to say "green" without even knowing why. But some proportion of the time we also claim, "I have a subjective visual experience. I *see* that thing with my conscious mind. Seeing *feels* like something." The same kind of subjective experience can pertain to other sensory events—a sound, a touch, heat and cold, and so on.

Consider another domain of information: episodic memory. It is a part of our self-identity. It provides a sense of a trajectory through life. But memory itself is not fundamentally mysterious. A computer can store memory, including elaborate information about its past states. Those memories can be retrieved and used to guide output. The crucial, human difference is not that we have memories, or that we can recall them, but that we have a subjective *experience* of memories as we recall them.

Consider one more information-processing event: a decision. Once more, decision-making is not fundamentally mysterious. A computer can make a decision. It can take in information, integrate it, and use it to select one course of action out of many. The human brain also makes decisions. Most of those decisions, possibly tens of thousands a day, occur automatically with no subjective experience, much like in a computer. Yet in some instances, we also report a subjective awareness of making the decision. We sometimes call it intention, choice, or free will. The ability to make a decision, in itself, is not a special human capability. The crucial difference between a personal computer and a human brain lies in the subjective experience that is, sometimes, associated with decision-making—or with memory, sensory processing, or other events in the brain.

This subjective experience is often called consciousness. I admit the term can be misleading. To some people, consciousness refers to a metaphysical soul that floats free of the body after death. To many people it refers to the rich contents swirling within a mind. To some it refers specifically to the part inside you that has free will and chooses one action over another. I mean none of these things. I am referring to the human claim that we have a subjective experience of anything at all. In this account, I will use the terms consciousness, subjective awareness, and subjective experience interchangeably, to refer to this phenomenological property that people claim is associated with some select events and information in the brain.

Like many scientists who study consciousness, I focus on a microcosmic problem: a person looking at a small round spot on a screen (e.g., Webb et al., 2016a). In some circumstances, the person could say, "I have a subjective experience of seeing that spot." In other circumstances, the spot is processed by the visual system, has a measurable impact on the person, and even affects the person's speech and decisions, and yet the person will report, "I didn't consciously see anything." What is the difference between these two circumstances? Why is subjective awareness attached to the visual event in one case and not the other? If we can understand the relevant brain processes for awareness of a spot on a screen, then in principle we can extend the explanation to any information domain. We would understand how people have a subjective experience of vision, touch, sound, the internal richness of memory, mental imagery, decision-making, and self. We would understand the conscious mind.

My point here is that most of what composes the conscious mind is, in principle, not a fundamental mystery. What has resisted explanation thus far is not the content of our experience, but the presence of subjective experience itself. I argue that subjective experience is a confined, relatively easy piece of the neural puzzle to solve.

I also argue that the solution is no mere philosophical flourish. Instead, it is a crucial part of the way the system models and controls itself. It is a key part of the engineering. Without understanding the subjective awareness piece, it may be impossible to build artificial intelligence that has a human-like ability to focus its computational resources and intelligently control that focus. It may also be impossible to build artificial intelligence that can interact with people in a socially competent manner. The study of consciousness is sometimes mistaken as a pursuit of metaphysical mystery, without any practical consequences. The AST does not address a metaphysical mystery. It addresses a concrete piece of the neural puzzle, as pragmatic as the transmission mechanism in a car.

GRASPING AN APPLE WITH THE HAND

The idea of an attention schema was developed in analogy to the body schema. The body schema is an internal model, a rich and integrated set of information that reflects the state of the body, how it moves, and its relationship to the world (Head and Holmes, 1911; Shadmehr and Mussa-Ivaldi, 1994; Graziano et al., 2000; Graziano and Botvinick, 2002; Holmes and Spence, 2004). The body schema not only contributes to the brain's control of the body but also contributes to cognition and verbal behavior. It allows the brain to draw conclusions and make claims about the body. Without a body schema, we would not know that we have a body—except in an intellectual sense, the same way we all know that we have a pancreas. With a body schema, we report having whatever shape or type of body is represented by that body schema. The present section describes the body schema and some of its implications. The following section will draw parallels to an attention schema and our claim to have awareness.

To understand the body schema, consider the body as a robotic device (it could be legitimately called a biological robot) and the brain as the information processor that controls it. Suppose this robot has reached out and grasped an apple. We want to know what information is available to that robot's brain. Three specific types of information are relevant to this discussion: information about the apple, about the robot's own body, and about the physical relationship between the robot and the apple. One of the most important and overlooked aspects of the body schema is that it is not just a representation of the body itself. It contains information about the relationship between the body and the rest of the world.

We will begin with the apple. We ask this biological robot what it is holding, and the robot answers, "An apple." We ask the robot, "Can you describe the apple?" and the robot does so. How does the robot do this? Its brain contains linguistic and cognitive machinery. The cognitive machinery has partial access to the models constructed within its visual system. Its visual system has constructed a rich model of the apple, a set of information about size, color, shape, location, and other attributes, constantly updated as new signals are processed. Due to the presence of this information, and due to the cognitive and linguistic access to the information, the machine is able to respond. It is worth noting that the robot is not actually telling you about the apple. It is telling you about the model of an apple, essentially a simulation, constructed in its visual system. If the internal model contains an error, if it represents the apple as twice too big, for example, the machine will report that incorrect information.

Next, we ask the robot, "What is the state of your body?" Once again, the robot can answer. The reason is that the brain has constructed a body schema—a set of information, constantly updated as new signals are processed, that specifies the size and shape of the limbs and torso and head, how they are hinged, the state they are in at each moment, and what state they are likely to be in over the next few moments. The primary purpose of a body schema is to allow the brain to control movement. A secondary consequence of the body schema is that the robot can explicitly talk about its body. Its cognitive and linguistic processors have some access to the body schema, and therefore the robot can describe its physical self.

Once again, it is worth noting that the robot is not reporting on the actual state of its body, but rather reporting the contents of an internal model. If that internal model is in error, then the robot will provide an incorrect report. If you trick the body schema into representing the arm as more to the left than it actually is, or larger than it actually is, that distorted information will pass through cognition and linguistic processing and enter the verbal report. Even rather extreme illusions of the body schema are easily induced, such as the rubber hand illusion (Botvinick and Cohen, 1998) or the Pinocchio illusion (Lackner, 1988). It is also worth noting that even when the body schema is working correctly, it is always incomplete. It does not contain information about, for example, bone structure, tendon attachments, or the biophysics of muscle contraction. Our biological robot cannot access its body schema and on that basis tell you about the actin and myosin fibers in the muscles. Its body schema contains only the information that the system needs to control the body. The body schema is, in a sense, a cartoon sketch of the body.

Finally, we ask the robot, "What is your physical relationship to the apple?" The robot says, "My arm is outstretched and my hand is grasping the apple." The answer requires integrating two different internal models: the visual system's model of the apple and the body schema. The machine has constructed an amazingly complex, brain-spanning meta-model. Yet in its essence, the behavior remains simple. The machine constructs internal models descriptive of its world. It can report the information content of those internal models because its cognitive and linguistic mechanisms have at least partial access to those internal models. Nothing here is mysterious. Nothing is outside the realm of engineering. I argue that the biological robot, as described thus far, could be copied in artificial form using today's engineering expertise, and it would function in essentially the same way.

I use the term "robot" to communicate a mechanistic perspective, but I intend to describe a human being. We operate in the manner described above. If you hold an apple, the reason why you can say so is that your brain has constructed an internal model of the apple and of your body, integrated those two models to form a larger, overarching description of your physical relationship to that apple, and cognitive and linguistic machinery has access to those internal models. There is something tautological about my central assertion: every claim a person makes, even a simple claim like, "Right now I'm holding an apple," depends on information constructed in the brain. Without the requisite information, the system would be unable to make the claim.

GRASPING AN APPLE WITH THE MIND

Suppose the robot as described above is asked another question. We ask it, "What is the mental relationship between yourself and the apple?" If the robot contains only an internal model of the apple and of a body schema, I argue that it would not be able to answer the new question. It would lack sufficient information. It has sufficient information to answer basic questions about its physical body, about the apple, and about the physical relationship between the two. But a mental relationship? It lacks information on what a mental relationship is. We could ask, "Are you conscious of the apple?" but given the information present, the machine could provide only concrete and literal information such as, "There is an apple." We could press and say, "Yes, but do you have an internal, subjective experience of it?" How could the machine answer? Thus far, we have not given it information to process that question. It would be like asking a digital camera whether it is aware of the picture it just took. The question is meaningless.

Almost all theories of consciousness focus on how a brain might generate a feeling of consciousness. The AST takes a more pragmatic approach, asking how a machine can make the claim that it has a subjective experience. It is a theory about how the brain constructs the requisite information such that the person can make that specific claim. Without the requisite information, the claim cannot be made.

The AST is, in a sense, a proposed extension of the body schema. The proposal is that the brain constructs not only a model of the physical body but also a model of its own internal, information-handling processes. It constructs an "attention schema." That attention schema not only contributes to the control of attention but the information contained within it also has consequences for the kinds of claims that the machine can make about itself.

Attention is a catchall term that arguably adds more confusion than clarity, given its many connotations and meanings. Here, I will mainly avoid the term and use the phrase, "enhanced processing." I will occasionally use the term "attention" when nothing else captures the intended meaning succinctly. The phenomenon I outline below matches at least some uses of the term attention, especially as described by the neuroscientific, "biased competition" theory of attention (Desimone and Duncan, 1995; Beck and Kastner, 2009).

Signals in the brain can be selectively enhanced. For example, consider again the robot from the previous section that encounters an apple. Its visual system constructs a representation of the apple. Under some circumstances, that representation may be suppressed in favor of other representations. Perhaps a sandwich, or another person, or something startling like a bear, wins a competition of visual signals, rises in signal strength, and suppresses the representation of the apple. Under other circumstances, the apple becomes the focus of processing and its representation is enhanced at the expense of other visual representations. This constantly shifting competition among signals can be slanted or biased toward one item or another by a variety of influences, including bottom-up influences (such as a suddenly moving object that causes a surge of signal in the visual system) or top-down influences (such as a cognitive decision to focus one's resources on a specific task). If the apple's representation in the visual system gains in signal strength, winning the competition of the moment, that enhanced processing has a suite of consequences. The apple is processed in greater depth—its nuances and details are more fully processed. It is also more likely to affect other systems throughout the brain, beyond the visual system. The signal is, in effect, broadcasted to other brain areas. It is therefore more likely to affect behavioral decision-making. Whether you reach for the apple or not, bite it, put it away, or decide not to touch it because it looks rotten, the processing of the apple has an impact on behavioral choice. The apple is also more likely to impact memory, allowing it to be recalled later and affect future behavior.

The focusing of resources described here is not limited to a spatial focus. One can focus processing resources on color, on motion, on a particular shape, or on other non-spatial features. It is also not limited to vision. The same type of selective, enhanced processing can be seen in audition, touch, and presumably smell and taste. One can apply the same enhanced processing to movement commands during a difficult movement sequence. It is even possible to selectively enhance entirely internal signals, such as recalled memories, visual imagination, or internal speech. The constantly shifting, enhanced processing of some signals over others, across a vast range of information domains, is one of the most fundamental attributes of the brain.

Now consider again the robot holding an apple. Suppose the machine is focusing its processing resources on the apple. You ask the robot, "What is your mental relationship to the apple?" Can the robot answer this question? Does it have sufficient internal information to report what it is doing computationally? According to AST, the robot can indeed answer the question, and the reason is that it contains an attention schema. The attention schema is a set of information that describes the act of focusing resources on something. The attention schema describes what attention is, what it does, what its most basic stable properties are, what its dynamics and consequences are, and monitors its constantly changing state. Given the information in the attention schema, and given cognitive and linguistic access to at least some of that information, the machine is able to say, "I have a mental grasp of the apple."

Just as the body schema lacks information about mechanistic details such as bone structure and tendon insertion points, so the proposed attention schema lacks detailed information about how signals in the brain are selectively enhanced. The proposed attention schema lacks information about neurons, synapses, electrochemical signals, neural competition, and so on. It has a relatively impoverished description. Suppose you ask the machine, "Tell me more about this mental possession. What physical properties does it have?" The machine is not going to be able to give a scientifically accurate answer. It cannot describe the neuroscience of attention. It replies on the basis of the information available in the attention schema. It says, "My mental possession of that apple, the mental possession in and of itself, has no describable physical properties. It just is. It's a non-physical part of me. My arms and legs are physical parts of me; they have substance. Whatever's inside me that has mental possession of things, that part is non-physical. It's metaphysical. It's my awareness."

It is important to point out what I am not saying. It is easy to imagine building a machine that says, "I am aware of the apple." Just record that message on your phone, then press play, and the machine will utter the phrase. That superficial solution is not what is being described here. What is crucial here is the presence of a rich, descriptive model that is constructed beneath the level of cognition and language, and yet still is accessible to cognition. Because the machine is responding on the basis of an internal model, the response can be flexible, self-consistent, and meaningful. If you ask the machine for more details, it can give a rich description. It might add, "That non-physical, subjective part of me, the real me, is located inside my body. It hovers in my head. It's more or less vivid depending on circumstances. Now that I'm aware of that apple, I *know* about it, what it is and what it's good for. I can choose to react to it. I'll be able to remember it for later. Those are just some of the consequences of awareness. And awareness is not limited to apples. I sometimes experience other things as well. Right now I'm aware of you, sometimes I experience a flood of recalled memories, or mental imagery that I invent fancifully, and sometimes I have the subjective experience of making a decision. There's a commonality across all those circumstances—I have a subjective, mental possession of things inside me and around me." In this description, the machine is coming close to the literal truth. It is giving a fairly close, if high-level and detail-poor, description of how it focuses its processing resources on one or another item. Its description veers from literal reality only as it muddles the more mechanistic details and ultimately claims to have a spooky, physically incoherent consciousness. Consciousness is, in a sense, a cartoon sketch of attention.

Suppose you ask the machine, "But aren't you making all those claims simply because that's the information contained in your internal models? Aren't you just a computing machine?"

The machine accesses its internal models and finds nothing to match your suggestion. Its internal models do not announce to cognition, "By the way, this is information contained in an internal model, and the information might not be literally accurate." On the basis of the limited information available, the machine says, "What information? What internal models? This has nothing to do with computation. No, I am simply subjectively aware of the apple." The machine is captive to its own information. It knows only what it knows.

Colleagues have often asked me: granted that the brain probably does construct something like an attention schema, how does that internal model explain how we have subjective experience? Why does it *feel* like anything at all to process information? The answer is that the theory emphatically does not explain how we have a subjective experience. It explains how a machine *claims* to have a subjective experience, and how it is that the machine cannot tell the difference.

The AST has some similarities to the illusionist approach to consciousness (e.g., Dennett, 1991; Norretranders, 1999; Frankish, 2016). In that view, subjective experience is not truly present; instead, the brain is an entirely mechanistic processor of information that has an illusion of possessing consciousness. Exactly how the illusion occurs differs somewhat between accounts. Clearly, the illusionist approach has a philosophical similarity to the AST. However, I remain uncomfortable with calling consciousness an illusion. In AST, the brain does not experience an illusion. It does not subjectively experience anything. Instead, the machine has wrong, or simplified information that tells it that it is having an experience. In my view, calling consciousness an illusion is trying too hard to employ an everyday, intuitive concept that is not truly applicable.

Another similar approach to consciousness might be called the "naïve theory" perspective (e.g., Gazzaniga, 1970; Nisbett and Wilson, 1977; Dennett, 1991). In that view, the brain processes information about its world but does not possess any subjective experience. We claim that we do because, at a cognitive level, we have learned a naïve theory. It is essentially a ghost story, a socially learned narrative that we use to explain ourselves, a social epiphenomenon with debatable utility. With different upbringing, we would not claim to have any conscious experience. Again, there is some philosophical similarity between this view and AST. Indeed, the two are very close. However, in AST, the naïve construct of consciousness is not learned. It is not at a higher cognitive level. It is wired into the system at a deep level and constructed automatically, like the body schema. It is inborn. As discussed below, it is probably present in a range of species. Moreover, it is not a social epiphenomenon; instead, it serves a specific set of important cognitive functions. The brain constructs internal models because of the specific usefulness of modeling and monitoring items in the real world, and the usefulness of the attention schema is the crux of the theory, as discussed in the following sections.

The AST also has strong similarities to approaches in machine consciousness (e.g., Chella et al., 2008) in which a system can contain representations of the self, the environment, and higher order, recursive representations of how the self relates to the environment. This general concept resonates closely with the concepts of the AST. The AST is a theory of how the human brain models its own human-like attention systems and thus makes the claim that it has a subjective experiential component. Artificial systems that have different internal architecture, perhaps different processes akin to but not identical to human attention, might require different self-representations. A machine of that nature would not necessarily lay claim to consciousness in the sense that we humans intuitively understand it. Drawing on its own internal quirky representations, it would describe itself in ways specific to it. Of course, we might expect the contents of that machine's mind to differ from a human's mind. But, the point I am trying to make here is that the very construct of consciousness, of subjective experience itself, whether the machine even has that construct and what the details of it may be, will depend on the precise nature of the machine's internal models.

THE ADAPTIVE VALUE OF AN ATTENTION SCHEMA: CONTROL OF ATTENTION

The sections above discussed the consequences of cognitive and verbal access to internal models. For example, the body schema allows you to close your eyes and still know about and talk about the configuration of your body. The primary function of the body schema, however, is probably less for cognitive access and more for the control of movement. One of the fundamental principles in control engineering is that a good controller contains a model of the item being controlled (Conant and Ashby, 1970; Francis and Wonham, 1976; Camacho and Bordons Alba, 2004; Haith and Krakauer, 2013). A robot arm, the airflow throughout a building, a self-driving car, each system benefits from an appropriate internal model. The model partly monitors the state of the item to be controlled and also partly predicts states into the near future. The body schema contains layers of information about the body, about its stable properties such as its shape and hinged structure and about more dynamic properties such as forces and velocities (Head and Holmes, 1911; Shadmehr and Mussa-Ivaldi, 1994; Shadmehr and Moussavi, 2000; Graziano and Botvinick, 2002; Holmes and Spence, 2004; Hwang and Shadmehr, 2005). This information is used during the control of movement for obstacle avoidance, for on-line error correction, and for longer term adaptation. If movements are systematically wrong or distorted, the internal model can be adapted to correct the errors.

We hypothesized that the same advantages accrue from having an attention schema. The ability to focus processing resources strategically on one or another signal requires control. That control should benefit from an attention schema—a coherent set of information that represents basic stable properties of attention, reflects ongoing changes in the state of attention, makes predictions about where attention can be usefully directed, and anticipates consequences of attention. The best way to test this hypothesis would be to isolate cases where awareness fails—cases where the brain is processing information but people report being unaware of it. In those cases, by hypothesis, the attention schema has failed. While the system may still be capable of directing attention, focusing resources on the signal in question, the control of attention should suffer in characteristic ways—much like the control of the arm might become more wobbly, less able to error-correct, and less adaptable over repeated trials, if the arm's internal model is compromised.

Several experimental results on attention and awareness have been interpreted as consistent with this prediction (McCormick, 1997; Tsushima et al., 2006; Lin and Murray, 2015; Webb and Graziano, 2015; Webb et al., 2016a), though more experiments are needed. Thus far, the relevant experiments have focused on visual attention and visual awareness. When people are unaware of a visual stimulus, they can still sometimes focus processing resources on it. They can direct attention to it (McCormick, 1997; Lamme, 2003; Woodman and Luck, 2003; Ansorge and Heumann, 2006; Tsushima et al., 2006; Kentridge et al., 2008; Hsieh et al., 2011; Norman et al., 2013). However, in that case, visual attention suffers deficits in control. It behaves less stably over time and shows evidence of being less able to error-correct and less able to adapt to perturbations (McCormick, 1997; Lin and Murray, 2015; Webb and Graziano, 2015; Webb et al., 2016a). The evidence suggests that awareness is necessary for the good control of attention.

One group of researchers has presented a computational model of attention with and without an internal model and found that at least this simplified, artificial attention is better controlled with the internal model (van den Boogaard et al., 2017).

In our hypothesis, the attention schema first evolved as a crucial part of the control system for attention. The possible co-evolution of attention and awareness has been discussed before (Graziano, 2010, 2013, 2014; Haladjian and Montemayor, 2015; Graziano and Webb, 2016). Since the basic vertebrate brain mechanisms for controlling attention emerged more than half a billion years ago, we speculate that the origin of awareness, at least in preliminary form, may be equally ancient. Awareness, in this view, is not simply a philosophical flourish. It is a part of the engineering. Just as one cannot understand how the brain controls the body without understanding that the brain constructs a body schema, so one cannot understand how the brain intelligently deploys its limited processing resources without understanding that it constructs an attention schema. That an attention schema causes us humans to lay claim to a metaphysical soul is a quirky side effect.

THE ADAPTIVE VALUE OF AN ATTENTION SCHEMA: SOCIAL COGNITION

One of the most devastating impairments to awareness in the clinical literature is hemispatial neglect. Damage to one side of the brain, typically the right temporoparietal junction (TPJ), causes a loss of awareness of everything to the opposite side of space (Vallar and Perani, 1986; Corbetta, 2014). Yet, information from the neglected side is still processed to some degree (Marshall and Halligan, 1988), and the visual system is still active to the highest levels of processing (Rees et al., 2000; Vuilleumier et al., 2002). Neglect appears to be caused by the disruption of brain networks involved in attention and awareness that pass through the TPJ (Corbetta, 2014; Igelström and Graziano, 2017).

The TPJ, however, has also been implicated in social cognition. When people attribute mind states to each other, such as beliefs or emotions, brain-wide networks are recruited that also pass through the TPJ (Saxe and Wexler, 2005; Kelly et al., 2014; Igelström et al., 2016). A complicated literature suggests that, although there is some separation of function among subregions of the TPJ, considerable overlap of function is also present (Mitchell, 2008; Scholz et al., 2009; Igelström et al., 2016; Igelström and Graziano, 2017). The adjacency and possible overlap of social cognition functions with awareness and attention functions has caused some controversy.

We suggested that the functional overlap within the TPJ may have a deeper significance (Graziano and Kastner, 2011; Graziano, 2013). In our proposal, one of the primary uses for the construct of awareness is for social cognition. We attribute to other people an awareness of the objects and events around them. When we do so, we are in effect constructing a simplified model of other people's state of attention. Arguably, all of social cognition depends on attributing awareness to other people. Does Frank intend to walk toward you, or sit in that chair, or eat that sandwich? Only if he is aware of you, the chair, or the sandwich. Is he angry that someone made a rude gesture at him? Only if he is aware of the gesture. Whether reconstructing someone else's beliefs, intentions, emotions, or any other mental state, we depend first on attributions of awareness.

In our hypothesis, the TPJ is a central node in a brain-wide network that helps to compute an attention schema. That attention schema is our construct of awareness, and that construct can be applied to oneself or to others. Much like the color-processing networks in the visual system can assigned colors to surfaces, so the social cognition network can assign the construct of awareness to agents, including oneself. Experimental evidence from brain imaging studies suggests that the TPJ does play a role in attributing visual awareness to others, and that some of the same subregions of the TPJ are involved in constructing one's own visual awareness (Kelly et al., 2014; Igelström et al., 2016; Webb et al., 2016b). We suggest that the TPJ is a site where the ability to perceive consciousness in others grew out of our ability to be conscious ourselves. However, the TPJ remains an extremely complex area of the cortex that is still poorly understood. Far more work will be needed to specify its range of functions and how they are distributed anatomically.

Given the goal of this article, introducing AST to those who may be interested in engineering it, the specific networks in the brain are not of great importance. Whether the computations are performed by this or that part of the brain are irrelevant. What is important is the overlap in function between modeling oneself and modeling others. A mechanism that can compute an internal model of attention, an attention schema, may be important not just for controlling one's own attention, but also for monitoring the attentional states of others. The social use of an attention schema may be especially developed in humans. We attribute awareness to each other, to pets, to inanimate objects, and to the spaces around us. Arguably, the entire spirit world, from deities down to minor ghosts, owes itself to our social neural machinery building the construct of awareness and attributing it promiscuously to ourselves and everything else around us. To build machines with similar social ability, the ability to attribute consciousness to itself and to others, such that the machine can understand what it means for another agent to be conscious, may require something like an attention schema.

WHY BUILD ARTIFICIAL CONSCIOUSNESS?

If AST is correct, then consciousness is buildable with current technology. In this respect, the theory differs from other major theories of consciousness that provide much less clear direction for how to build consciousness.

For example, the global workspace theory posits that the brain-wide boosting and broadcasting of a signal, such as a visual signal, causes that signal to enter consciousness (Baars, 1988; Dehaene, 2014). In effect, the global workspace theory is the same as the AST, if you took away the attention schema part, and had only the attention part—the ability of the brain to selectively enhance signals such that they have a global impact on many brain systems. While in my view the theory is likely to be correct as far as it goes, it is incomplete. It does not explain why the globally broadcasted information would be associated with the property of subjective experience. Building a machine that has signals boosted in that manner, to a strength sufficient to globally effect other systems in the machine, is easily done and arguably has already been done. But it is not a good prescription for building consciousness. There is no reason to suppose that a machine of that sort would sit up and say, "Wow, I have an internal experience of these things." It brings us no closer to the behavior that humans exhibit, namely, claiming to have subjective awareness.

The integrated information theory (Tononi, 2008) suffers a similar problem. In that theory, consciousness is the result of highly integrated information in the brain. A mathematical formula can tell you how much integrated information, and thus how much consciousness, is present in any specific device. To many scientists, including myself, this theory is non-explanatory and ultimately unfalsifiable. It is somewhat like the science fiction trope: if you build a computer big and complex enough, integrating enough information together, it will somehow become conscious. To be fair to the theory, in my view, there is likely to be at least some type of relationship between consciousness and highly integrated information. Even in AST, the proposed attention schema is a bundle of information that is integrated with other schemas and models around the brain. But as a prescription for building consciousness, the integrated information theory by itself has been disappointing, since even very complex technology that contains a lot of integrated information has not announced its consciousness yet.

The AST instead presents an extremely simple conceptual foundation. The machine claims to be conscious of items and events, because it constructs information that describes that condition of consciousness. Without the internal information indicating that it contains consciousness, it would not be able to make the claim. The reason why it constructs that quirky internal information is because it is a useful, if not literally accurate, model of the machine's ability for deep, focused processing. The AST therefore points a practical way toward building a machine that makes the same claims of consciousness that people do.

I recognize that AST is not yet specific enough to hand a blueprint to an engineer. Yet, it lays a conceptual foundation for building consciousness. Because it is a theory in which a machine constructs a specific set of information and uses it in a specific way, it is buildable. Given current technology, an enterprising set of AI researchers should be able to build a machine that contains a fairly rich model of what consciousness is and that can attribute the property of consciousness to itself and to the people it interacts with. It should be possible to build a machine that believes it is conscious and claims it is conscious and acts like it is conscious and that talks about its consciousness in the same ways that the human machine does.

Why try to build artificial consciousness? One could build it for entertainment value. It would be monumentally cool. But I also see two practical reasons. The first may be of technical interest to specialists, whereas the second is of fundamental importance to all of us.

First, evolution has given us effective brains, and copying the biological solution might make for capable artificial intelligence. Suppose that the theory is correct, and consciousness depends on an attention schema. With an attention schema acting as an internal control model, the brain is better able to control and deploy its limited processing resources. Perhaps giving machines a human-like focus of attention, and an attention schema, will be helpful. Artificial systems might thereby become better able to control their own limited processing resources. Admittedly, I do not know if this engineering trick borrowed from the brain will be of use to artificial intelligence. Computer systems can process more information, more quickly, than biological systems, and can be organized in fundamentally different ways. It is not clear whether human-like attention, or human-like control of attention, would necessarily benefit artificial systems. The idea would be worth pursuing, but better engineering solutions might be discovered along the way.

To me the most compelling reason to pursue artificial consciousness is that, if the theory is correct, then consciousness is the foundation of social intelligence. An agent cannot be socially competent unless it has a fairly rich internal model of what consciousness is and can attribute consciousness to itself and to other people. If we want to build machines that are skilled at interacting with people, we will need to build in consciousness in the same sense that people attribute consciousness to themselves and see consciousness in others. It is the root of empathy. Without that capacity, our computers are sociopaths. A similar point has been made by others, including the point that social capability is urgently needed in artificial intelligence (e.g., Sullins, 2016), and that self-models are a crucial part of human social competence (e.g., Hood, 2012).

While human sociopaths are evidently conscious—they can attribute that property to themselves—they are impaired at attributing it to others. They may know intellectually that other people contain minds, but they appear to lack a fundamental, automatic perception of the consciousness of others. Other people are mechanical objects to them. Half of the functional range of the attention schema is impaired. We cannot build machines that treat people with humanistic care, if they do not have that crucial social capability to attribute consciousness to others. Machine consciousness is a necessary step for our future. For those who fear that AI is potentially dangerous and may harm humanity, I would say that the danger is infinitely greater with sociopathic computers and it is of the utmost priority to give them consciousness—both the ability to attribute it to themselves and to others. I urge anyone with the technical expertise, who is reading this article, to think about how to tackle the problem.

AUTHOR CONTRIBUTIONS

MG is responsible for all aspects of this article.

FUNDING

This study was funded by the Princeton Neuroscience Institute Innovation Fund.

REFERENCES

Ansorge, U., and Heumann, M. (2006). Shifts of visuospatial attention to invisible (metacontrast-masked) singletons: clues from reaction times and event-related potentials. *Adv. Cogn. Psychol.* 2, 61–76. doi:10.2478/v10053-008-0045-9

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

Beck, D. M., and Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vis. Res.* 49, 1154–1165. doi:10.1016/j.visres.2008.07.012

Botvinick, M., and Cohen, J. D. (1998). Rubber hand 'feels' what eye sees. *Nature* 391, 756. doi:10.1038/35784

Camacho, E. F., and Bordons Alba, C. (2004). *Model Predictive Control*. New York: Springer.

Chella, A., Frixione, M., and Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artif. Intell. Med.* 44, 147–154. doi:10.1016/j.artmed.2008.07.003

Conant, R. C., and Ashby, W. R. (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

Corbetta, M. (2014). Hemispatial neglect: clinic, pathogenesis, and treatment. *Semin. Neurol.* 34, 514–523. doi:10.1055/s-0034-1396005

Dehaene, S. (2014). *Consciousness and the Brain*. New York: Viking.

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

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi:10.1146/annurev.ne.18.030195.001205

Francis, B. A., and Wonham, W. M. (1976). The internal model principle of control theory. *Automatica* 12, 457–465. doi:10.1016/0005-1098(76)90006-6

Frankish, K. (2016). Illusionism as a theory of consciousness. *J. Conscious. Stud.* 23, 11–39.

Gazzaniga, M. S. (1970). *The Bisected Brain*. New York: Appleton Century Crofts.

Graziano, M. S. A. (2010). *God, Soul, Mind, Brain: A Neuroscientists Reflections on the Spirit World*. Fredonia: Leapfrog Press.

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. New York: Oxford University Press.

Graziano, M. S. A. (2014). Speculations on the evolution of awareness. *J. Cogn. Neurosci.* 26, 1300–1304. doi:10.1162/jocn_a_00623

Graziano, M. S. A., and Botvinick, M. M. (2002). "How the brain represents the body: insights from neurophysiology and psychology," in *Common Mechanisms in Perception and Action: Attention and Performance XIX*, eds W. Prinz and B. Hommel (Oxford: Oxford University Press), 136–157.

Graziano, M. S. A., Cooke, D. F., and Taylor, C. S. R. (2000). Coding the location of the arm by sight. *Science* 290, 1782–1786. doi:10.1126/science.290.5497.1782

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. (2016). "From sponge to human: the evolution of consciousness," in *Evolution of Nervous Systems*, 2nd Edn, Vol. 3, ed. J. Kaas (Oxford: Elsevier), 547–554.

Haith, A. M., and Krakauer, J. W. (2013). "Model-based and model-free mechanisms of human motor learning," in *Progress in Motor Control: Advances in Experimental Medicine and Biology*, Vol. 782, eds M. Richardson, M. Riley, and K. Shockley (New York: Springer), 1–21.

Haladjian, H. H., and Montemayor, C. (2015). On the evolution of conscious attention. *Psychon. Bull. Rev.* 22, 595–613. doi:10.3758/s13423-014-0718-y

Head, H., and Holmes, G. (1911). Sensory disturbances from cerebral lesions. *Brain* 34, 102–254. doi:10.1093/brain/34.2-3.102

Holmes, N., and Spence, C. (2004). The body schema and the multisensory representation(s) of personal space. *Cogn. Process.* 5, 94–105. doi:10.1007/s10339-004-0013-3

Hood, B. (2012). *The Self Illusion: How the Social Brain Creates Identity*. New York: Oxford University Press.

Hsieh, P., Colas, J. T., and Kanwisher, N. (2011). Unconscious pop-out: attentional capture by unseen feature singletons only when top-down attention is available. *Psychol. Sci.* 22, 1220–1226. doi:10.1177/0956797611419302

Hwang, E. J., and Shadmehr, R. (2005). Internal models of limb dynamics and the encoding of limb state. *J. Neural Eng.* 2, S266–S278. doi:10.1088/1741-2560/2/3/S09

Igelström, K. M., and Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: a network perspective. *Neuropsychologia*. 105, 70–83. doi:10.1016/j.neuropsychologia.2017.01.001

Igelström, K., Webb, T. W., and Graziano, M. S. A. (2016). Functional connectivity between the temporoparietal cortex and cerebellum in autism spectrum disorder. *Cereb. Cortex* 27, 2617–2627. doi:10.1093/cercor/bhw079

Kelly, Y. T., Webb, T. W., Meier, J. D., Arcaro, M. J., and Graziano, M. S. A. (2014). Attributing awareness to oneself and to others. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5012–5017. doi:10.1073/pnas.1401201111

Kentridge, R. W., Nijboer, T. C., and Heywood, C. A. (2008). Attended but unseen: visual attention is not sufficient for visual awareness. *Neuropsychologia* 46, 864–869. doi:10.1016/j.neuropsychologia.2007.11.036

Lackner, J. R. (1988). Some proprioceptive influences on the perceptual representation of body shape and orientation. *Brain* 111, 281–297. doi:10.1093/brain/111.2.281

Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends Cogn. Sci.* 7, 12–18. doi:10.1016/S1364-6613(02)00013-X

Lin, Z., and Murray, S. O. (2015). More power to the unconscious: conscious, but not unconscious, exogenous attention requires location variation. *Psychol. Sci.* 26, 221–230. doi:10.1177/0956797614560770

Marshall, J. C., and Halligan, P. W. (1988). Blindsight and insight in visuo-spatial neglect. *Nature* 336, 766–767. doi:10.1038/336766a0

McCormick, P. A. (1997). Orienting attention without awareness. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 168–180. doi:10.1037/0096-1523.23.1.168

Mitchell, L. P. (2008). Activity in the right temporo-parietal junction is not selective for theory-of-mind. *Cereb. Cortex* 18, 262–271. doi:10.1093/cercor/bhm051

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know – verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi:10.1037/0033-295X.84.3.231

Norman, L. J., Heywood, C. A., and Kentridge, R. W. (2013). Object-based attention without awareness. *Psychol. Sci.* 24, 836–843. doi:10.1177/0956797612461449

Norretranders, T. (1999). *The User Illusion: Cutting Consciousness Down to Size*. New York: Penguin.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C., and Driver, J. (2000). Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain* 123, 1624–1633. doi:10.1093/brain/123.8.1624

Saxe, R., and Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43, 1391–1399. doi:10.1016/j.neuropsychologia.2005.02.013

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., and Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS ONE* 4:e4869. doi:10.1371/journal.pone.0004869

Shadmehr, R., and Moussavi, Z. M. (2000). Spatial generalization from learning dynamics of reaching movements. *J. Neurosci.* 20, 7807–7815.

Shadmehr, R., and Mussa-Ivaldi, F. A. (1994). Adaptive representation of dynamics during learning of a motor task. *J. Neurosci.* 14, 3208–3224.

Sullins, J. (2016). Artificial phronesis and the social robot. *Front. Artif. Intell. Appl.* 290:37–39. doi:10.3233/978-1-61499-708-5-37

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242. doi:10.2307/25470707

Tsushima, Y., Sasaki, Y., and Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science* 314, 1786–1788. doi:10.1126/science.1133197

Vallar, G., and Perani, D. (1986). The anatomy of unilateral neglect after right-hemisphere stroke lesions. A clinical/CT-scan correlation study in man. *Neuropsychologia* 24, 609–622. doi:10.1016/0028-3932(86)90001-1

van den Boogaard, E., Treur, J., and Turpijn, M. (2017). "A neurologically inspired neural network model for Graziano's attention schema theory for consciousness," in *International Work Conference on the Interplay between Natural and Artificial Computation: Natural and Artificial Computation for Biomedicine and Neuroscience, Part 1*, 10–21. doi:10.1007/978-3-319-59740-9_2

Vuilleumier, P., Armony, J. L., Clarke, K., Husain, M., Driver, J., and Dolan, R. J. (2002). Neural response to emotional faces with and without awareness: event-related fMRI in a parietal patient with visual extinction and spatial neglect. *Neuropsychologia* 40, 2156–2166. doi:10.1016/S0028-3932(02)00045-3

Webb, T. W., and Graziano, M. S. A. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Front. Psychol.* 6:500. doi:10.3389/fpsyg.2015.00500

Webb, T. W., Kean, H. H., and Graziano, M. S. A. (2016a). Effects of awareness on the control of attention. *J. Cogn. Neurosci.* 28, 842–851. doi:10.1162/jocn_a_00931

Webb, T. W., Igelström, K., Schurger, A., and Graziano, M. S. A. (2016b). Cortical networks involved in visual awareness independently of visual attention. *Proc. Natl. Acad. Sci. U.S.A.* 113, 13923–13928. doi:10.1073/pnas.1611505113

Woodman, G. F., and Luck, S. J. (2003). Dissociations among attention, perception, and awareness during object-substitution masking. *Psychol. Sci.* 14, 605–611. doi:10.1046/j.0956-7976.2003.psci_1472.x

Conflict of Interest Statement: 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.

ORIGINAL RESEARCH
published: 26 January 2018
doi: 10.3389/frobt.2018.00001

Humanoid Cognitive Robots That Learn by Imitating: Implications for Consciousness Studies

James A. Reggia[1,2]*, *Garrett E. Katz*[1] *and Gregory P. Davis*[1]

[1] Department of Computer Science, University of Maryland, College Park, MD, United States, [2] Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, United States

While the concept of a conscious machine is intriguing, producing such a machine remains controversial and challenging. Here, we describe how our work on creating a humanoid cognitive robot that learns to perform tasks *via* imitation learning relates to this issue. Our discussion is divided into three parts. First, we summarize our previous framework for advancing the understanding of the nature of phenomenal consciousness. This framework is based on identifying computational correlates of consciousness. Second, we describe a cognitive robotic system that we recently developed that learns to perform tasks by imitating human-provided demonstrations. This humanoid robot uses cause–effect reasoning to infer a demonstrator's intentions in performing a task, rather than just imitating the observed actions verbatim. In particular, its cognitive components center on top-down control of a working memory that retains the explanatory interpretations that the robot constructs during learning. Finally, we describe our ongoing work that is focused on converting our robot's imitation learning cognitive system into purely neurocomputational form, including both its low-level cognitive neuromotor components, its use of working memory, and its causal reasoning mechanisms. Based on our initial results, we argue that the top-down cognitive control of working memory, and in particular its gating mechanisms, is an important potential computational correlate of consciousness in humanoid robots. We conclude that developing high-level neurocognitive control systems for cognitive robots and using them to search for computational correlates of consciousness provides an important approach to advancing our understanding of consciousness, and that it provides a credible and achievable route to ultimately developing a phenomenally conscious machine.

Keywords: machine consciousness, artificial consciousness, neural network gating mechanisms, cognitive robots, cognitive phenomenology, imitation learning, computational explanatory gap, working memory

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo,
Italy

Reviewed by:
Owen E. Holland,
University of Essex, United Kingdom
Igor Aleksander,
Imperial College London,
United Kingdom

****Correspondence:***
James A. Reggia
reggia@cs.umd.edu

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *29 October 2017*
Accepted: *09 January 2018*
Published: *26 January 2018*

Citation:
Reggia JA, Katz GE and Davis GP (2018) Humanoid Cognitive Robots That Learn by Imitating: Implications for Consciousness Studies. Front. Robot. AI 5:1. doi: 10.3389/frobt.2018.00001

INTRODUCTION

In this paper, we use the word "consciousness" to mean specifically phenomenal consciousness unless explicitly indicated otherwise. The term "phenomenal consciousness" has been used historically to refer to the subjective qualities of sensory phenomena, emotions, and mental imagery, for example the color of a lemon or the pain associated with a toothache (Block, 1995). Searle has presented a list of essential/defining features of consciousness, including subjectivity, unity, qualitativeness, situatedness, and sense of self (Searle, 2004), and a detailed analysis of this term can be found in Chapter 3 in Tani (2017). Recent work in philosophy has argued for an extended view of phenomenology that includes one's cognitive processes and hence is referred to as cognitive phenomenology, as we will elaborate below. In the following, we focus on conscious qualities specific to cognitive phenomenology in particular, as opposed to the more historically emphasized aspects of consciousness such as sensory qualia.

How can research based on cognitive humanoid robots contribute to our understanding of consciousness? Consciousness is not well understood at present, and many philosophers have questioned whether computational studies or cognitive robots can play a significant role in understanding it. Such arguments cannot be refuted at present because there is currently no convincing implementation of instantiated consciousness in a machine, as described in Reggia (2013). Conversely, none of these past arguments appear sufficiently strong to convince many current investigators that machine consciousness is impossible (Reggia et al., 2015). For this reason, it seems prudent to us to push ahead investigating this issue until the matter can be definitively resolved one way or the other, and it is in that context that we describe our research efforts below.

Here, we describe how our past and ongoing work on creating a humanoid cognitive robot that learns to perform tasks *via* imitation learning relates to consciousness studies. Our key contribution here is to expand and develop a concrete framework for investigating the nature of consciousness in cognitive robots. Our discussion is divided into three parts. First, we summarize our framework for advancing the understanding of the nature of phenomenal consciousness based on studying the computational explanatory gap (CEG) (Reggia et al., 2014). The main goal in this work is to identify neurocomputational correlates of consciousness. We believe that identifying such correlates will be possible in cognitive robots, based on concepts that have emerged recently in the philosophical field of cognitive phenomenology, and we explain why that is so.

The core idea of our framework for studying consciousness in robots is that investigating how high-level cognitive processes are implemented *via* neural computations is likely to lead to the discovery of new computational correlates of consciousness. Accordingly, in the second part of this paper, we describe a cognitive robotic system that we recently developed that learns to perform tasks by imitating human-provided demonstrations. This humanoid robot uses cause–effect reasoning to infer a demonstrator's goals in performing a task, rather than just imitating the observed actions verbatim. Its cognitive components center on top-down control of a working memory that retains the explanatory interpretations that the robot constructs during learning. Because, as we explain below, both cause–effect reasoning and working memory are widely recognized to be important aspects of conscious human thought, we suggest that exploring how the cognitive and memory mechanisms embodied in our imitation learning robot provide an excellent test of our framework for studying consciousness in machines.

Finally, in the third part of this paper, we describe our recent and ongoing work that is focused on converting our robot's imitation learning cognitive system into purely neurocomputational form, including its causal reasoning mechanisms and cognitive control of working memory. We summarize our initial results exploring the feasibility of this idea. Based on these results, we argue that the top-down cognitive control of working memory, and specifically its gating mechanisms, is potentially an important computational correlate of consciousness in humanoid robots that merits much further study. We conclude that developing neurocognitive control systems for cognitive robots and using them to search for computational correlates of consciousness provides an important approach to advancing our understanding of consciousness, and that it provides a credible and achievable route to ultimately developing a phenomenally conscious machine.

A COMPUTATIONAL APPROACH TO UNDERSTANDING THE NATURE OF CONSCIOUSNESS

In the following, we propose a *computational* framework for investigating consciousness. We begin by summarizing the concept of a CEG, and we explain why recent advances by philosophers interested in cognitive phenomenology makes this barrier relevant to consciousness studies. We then describe our proposed framework for studying consciousness that is based on identifying its computational correlates.

Computation, Mind, Brain, and Body

We have previously suggested that there is an important obstacle to understanding the prospects for machine consciousness that we call the CEG (Reggia et al., 2014). The CEG is defined as our current inability to understand how higher-level cognitive computations supported by the brain can be accounted for by lower-level neurocomputational processes. We use the term "higher-level cognition" to refer to cognitive processes including decision-making, reasoning, intent-directed problem solving, executive control of working memory contents, plan generation, and language. These cognitive processes are viewed by many psychologists as being consciously accessible. In contrast, we use the term "lower-level neurocomputational processes" to refer to the types of computations that can be implemented using artificial neural networks like those currently studied in fields such as neuroscience, computer science, psychology, and engineering.

The CEG is related to past work in philosophy, neuroscience, and psychology, addressing various aspects of the mind–brain problem. In philosophy, the CEG differs from the *philosophical explanatory gap,* the latter referring to the difficulty we have in explaining how physical systems in the objective world can support the subjective qualities of consciousness (Levine, 1983). The philosophical explanatory gap relates to how difficult it is to understand how subjectivity can emerge from the brain or potentially from other physical systems such as machines. The CEG differs in that it is *not* a mind–brain issue. Instead, the CEG is our current inability to understand how *computations* supporting high-level cognitive processes like those described above can be implemented *via* the lower-level *computations* that neural networks provide. Put otherwise, it deals only with computational issues, and it applies both to people and to machines. Historically, philosophers have tended to deprecate the CEG, characterizing it as part of the "easy" problem of interpreting how the brain generates intelligent behavior (Chalmers, 1996). This viewpoint fails to account for why the CEG has been so difficult to bridge over the last 50 years in spite of an enormous research effort to do so. It also ignores the possibility that the philosophical explanatory gap and the CEG are not two independent issues, but that instead, the CEG might ultimately prove relevant to understanding the mind–brain problem. It is this latter issue that we discuss in the following, arguing that the CEG is relevant to obtaining a deeper understanding of the mind–brain problem. More recently in philosophy, work in *cognitive phenomenology* has argued that our phenomenal experiences are not limited to classical qualia such as those of sensory perception, but also include high-level cognition (Bayne and Montague, 2011; Jorba and Vincente, 2014; Chudnoff, 2015). It is this idea more than anything else that makes the CEG, a purely computational issue, of relevance to understanding consciousness. Accepting that some facets of cognition reach conscious awareness is what makes computational studies of the CEG important in consciousness studies. The *hypothesis* guiding our work described below is thus that bridging the CEG provides a pathway to deeper comprehension of consciousness and eventually possibly even a phenomenally conscious machine. This hypothesis makes research that is directed at creating neurocomputational implementations of higher-level cognitive processes, including our own work with adaptive cognitive robots as described below, relevant to the issue of phenomenal consciousness.

The CEG also relates to recent work in the neurosciences and psychology. In the neurosciences, our current state of knowledge can be characterized as knowing a lot about how high-level cognitive functions correlate with different macroscopic brain areas (e.g., language comprehension and Wernicke's area, planning and prefrontal cortex) and a great deal about the microscopic neurobiological networks in these same areas. However, what we do not currently understand is how the brain implements the high-level cognitive processes using the underlying neural circuitry. We view this situation as an example of the CEG, quite separate from any considerations about consciousness. In psychology, related work has been done to investigate the differences between information processing that is unconscious and information processing that is conscious (Dehaene and Naccache, 2001; Baars, 2002). Unconscious information processing is fast and can support multiple concurrent tasks, and these tasks can be done simultaneously without interfering with each other. It tends to involve localized brain regions and is often not reportable (people cannot explain how they carried out a task). In contrast, conscious information processing is much slower, restricted to one task at a time, involves widespread cortex activation, and is generally taken to be cognition that a subject can report. Again, we view such findings as being related to the CEG. The computational properties associated with unconscious processes often match up well with those of neural computations (e.g., the opaqueness or "non-reportability" of what a neural network has learned). The computational properties during conscious, reportable cognitive activities are much closer to what is seen with symbolic artificial intelligence (AI) systems, and do not relate well to how neural networks process information. To be clear, we are not suggesting that consciousness can be explained by symbolic reasoning or language—we just intend to convey that conscious, reportable cognitive activities need to be accounted for by resolving the CEG. Further, we are only considering the existence of consciousness in adults and do not relate our work to the mechanisms underlying the emergence of consciousness in infants.

Symbolic AI models are often used on computers devoid of any remotely human- or animal-like embodiment. However, all compelling and widely accepted examples of consciousness in the real world occur in embodied biological systems. Even proponents of cognitive phenomenology still consider it plausible that conscious cognitive processing has some basis in sensorimotor experience (Prinz, 2011). From a purely practical standpoint, studying the CEG in the context of embodied robotic systems may be the most efficient route to ecologically valid input data for cognitive models. And it stands to reason that humanoid robots in particular will be best for studying machine consciousness that is as human-like as possible. At a deeper level, there are serious philosophical positions that consider embodiment to be intrinsically related to cognitive phenomenology (Nagataki and Hirose, 2007). In sum, studying cognition in the context of humanoid robots specifically may be an important factor in bridging the CEG and potentially understanding/engineering consciousness.

A Framework for Investigating Consciousness

An implication of the ideas presented in the preceding section is that much recent research involving neurocomputational models of high-level cognition becomes relevant to comprehending the properties of consciousness. The basic idea is that these computational investigations could discover neurocomputational mechanisms occurring with phenomenally conscious aspects of cognition that are not also found to be present during cognitive processes that are unconscious. We have proposed elsewhere that this could provide examples of computational correlates of consciousness, in the same way that neuroscientists have identified neural correlates of consciousness (Reggia et al., 2014, 2016).

A *computational correlate of consciousness* has been defined previously to be an aspect of information processing associated with conscious but not unconscious information processing (Cleeremans, 2005). In general, a computational correlate of consciousness is not the same thing as a neural correlate as described by neuroscientists. Previously described neural correlates have included biological concepts that are not computational, e.g., regions of the brain, biochemical processes, and electrical activity patterns in the brain (Chalmers, 2000). On the other hand, the definition of computational correlates above is fairly general. For example, it might include logical reasoning algorithms like those studied in traditional AI. In this context, previous researchers have suggested that cognitive processes can be separated into neurocomputational processes representing unconscious facets of cognition, and symbolic processes representing conscious facets of cognition (Kitamura et al., 2000; Sun, 2002; Chella, 2007), i.e., symbolic information processing is viewed as a computational correlate of consciousness. However, from our perspective, such models do not provide a way to bridge the CEG. The central idea in bridging the CEG as we defined it above is to identify how higher-level reasoning is implemented *via* underlying, purely neurocomputational mechanisms, much as the brain does. This is the crux of the matter.

Thus, in the rest of this paper we use the term "computational correlates of consciousness" to refer solely to neurocomputational mechanisms that occur only with conscious facets of higher-level cognitive processes and are *not* found with neurocomputational processes involved with other unconscious information processing (not with neurocomputational mechanisms associated with implementing the normal pupil light reflex, for example). These correlates may be implemented in the brain, but are independent of the physical mechanisms that implement them (robot control circuitry, biological brain circuitry, and so forth). Our proposal is that uncovering computational correlates of consciousness will provide insight into the nature of consciousness (as per cognitive phenomenology) and possibly even the development of a plausibly conscious physical machine.

We have recently given a fairly detailed description of previously proposed computational correlates of consciousness (Reggia et al., 2016) and refer the interested reader to that work. Here, we just briefly give a few examples that illustrate the central ideas involved. One widely known proposal is that *global information processing* is a computational correlate of consciousness, inspired by findings that information processing during conscious mental activities (and not unconscious cognitive processes) occurs widely across the cerebral cortex and is also correlated with enhanced communication between brain regions (Baars et al., 2003; Massimini et al., 2005; Tagliazucchi et al., 2016). Another prominent past suggestion is that *information integration* in a neural network is what distinguishes conscious from unconscious systems in general (Tononi, 2004). Still others have suggested that having a *self-model* is a computational correlate (Searle, 2004; Samsonovich and Nadel, 2005), even showing that physical robots controlled by neural networks can pass the "mirror test" of self-awareness used with animals (Takeno, 2013). Other researchers have suggested that higher-order representations of one's knowledge about the world correlate with consciousness (Cleeremans et al., 2007; Pasquali et al., 2010). Additional studies have argued that *attention mechanisms* are potential computational correlates (Taylor, 2007; Haikonen, 2012). All of these ideas are intriguing and may provide important clues as to the fundamental nature of consciousness, and the fact that so many ideas are emerging in this area is quite encouraging.

A COGNITIVE HUMANOID ROBOT THAT LEARNS BY IMITATING

In the previous section, we described a framework for studying aspects of consciousness based on developing computational/robotic systems that account for high-level cognitive functions in neurocomputational terms. To pursue this approach, two things are needed: a physical robotic system that supports some aspects of high-level cognitive functionality, and an underlying neural control mechanism that implements that functionality.

Here, we describe our recent work on the first of these two requirements: Our efforts to create a cognitive humanoid robot that that can be used to explore consciousness-related and other issues (Katz et al., 2017a,b). Why would one want to consider studying the CEG in a robot instead of simply going the easier route of computer simulations? One answer is that a cognitive system in a robot is embodied: It interacts with and causally acts on a real external environment, and in that sense there is a true "mind-body" problem, at least to the extent that one is willing to call a robot's cognitive control system a mind. Further, it has been claimed that the ability to ground a cognitive robotic system's symbols in the robot's sensory data stream is a computational correlate of consciousness (Kuipers, 2008). While this suggestion is controversial (Chella and Gaglio, 2012), it suggests that some computational correlates may be particularly evident in a cognitive system that interacts with the real world as part of a physical system.

Our own robot learns to perform tasks by imitating human-provided demonstrations. During learning, it uses cause–effect reasoning to infer a demonstrator's goals in performing a task, rather than just imitating the observed actions literally. Importantly for our own research as described in subsequent sections, the robot's cognitive components center on top-down control of a working memory that retains the explanatory interpretations that the robot constructs during learning. We first briefly summarize this work here and then, in the next section, we relate this work to the search for computational correlates of consciousness.

Imitation Learning *via* Cause–Effect Reasoning

Our work in robotics is motivated in part by the fact that it is currently very hard to program humanoid robots to carry out multi-step tasks unless one has a great deal of expertise in robotics. A potential solution to this problem is to use imitation learning (learning from demonstrations) rather than manually programming a robot. With imitation learning, a robot watches a person perform the task to be learned, and then imitates what it observed. An important mode of imitation learning occurs at the sensorimotor level, when the learning robot closely imitates the motions, gestures, and perhaps even the facial expressions of the demonstrator. Much work on robotic imitation learning has focused on this level. While important, this level does not involve an understanding of the demonstrator's intentions, and hence suffers from limited ability to generalize to new situations where the robot must use different actions to carry out the same intentions.

Figuring out what a demonstrator's goals are is a kind of cause–effect reasoning known as "abduction" in AI. The issue is to postulate what the demonstrator's goals are in a way that is consistent with these goals *causing* the observed actions. AI researchers have extensively studied cause–effect reasoning (also called abductive reasoning) like this, including its use to infer the goals of an acting agent (Kautz and Allen, 1986; Peng and Reggia, 1990; Carberry, 2001). While some aspects of cognition have been simulated during past studies of imitation learning (Chella et al., 2006; Friesen and Rao, 2010; Dindo et al., 2011), to our knowledge, the utility of causal reasoning during imitation/

goal learning has not been studied substantially. However, in other application domains such as medical diagnosis or circuit fault localization, causal reasoning systems often rely on finite databases of background knowledge that exhaustively describe all of the possible causal events that might occur. In robotic imitation learning, this amounts to a finite list of general purpose primitive actions that a demonstrator or robot might perform, as well as the direct causal relationships between those actions and higher-level goals, and the possible objects that might be present in the environment. The full spectrum of possible goals, actions, and objects involved in general human imitation learning is probably too rich and variable to be adequately encoded in a finite database. However, for specific applications, there will likely be a finite set of possible objects to be manipulated and a finite set of actions and goals that can be applied to those objects. In this case, it is feasible to adapt existing causal reasoning approaches to robotic imitation learning. Moreover, individual actions and goals within a finite list can still admit continuous-valued parameters, such as object positions and rotations, in order to approximate some of the richness and variability inherent in true human imitation learning. This is the causal knowledge representation supported in our existing work described below. A detailed description of the encoded knowledge as well as the algorithms used in our applications can be found in Katz et al. (2017a). Future work on underlying neural mechanisms for the causal reasoning functionality could incorporate generative neural models to produce novel situation-specific actions that need not be anticipated in a finite database by a human knowledge engineer.

In this context, we recently suggested that causal reasoning is an important part of cognitively oriented imitation learning. To examine whether this idea can support imitation learning, we developed and studied an approach to imitation learning based on abductive cause–effect reasoning as illustrated in **Figure 1** (Katz et al., 2016, 2017a). During the observation of a demonstration, our approach assembles a parsimonious explanation for what was observed where the demonstrator's intentions (goals) serve to explain the actions performed by the demonstrator. We refer to our cognitive learning model as CERIL, for Cause–Effect Reasoning in Imitation Learning. The basic idea with CERIL is that the inferred demonstrator's goals (rather than the specific actions the demonstrator performed) can subsequently be used in related but new situations that may need different specific action sequences to achieve the same goals. Given that our primary interest here is in the role played by high-level cognition during imitation learning, our focus is on that and we largely take low-level sensorimotor processing as a given.

Figure 1 illustrates an example of CERIL learning about and then subsequently performing actions on a disk drive docking station. CERIL learns to maintain this disk drive dock, for example replacing hard drives that experience a hardware fault. The objective of learning is to replicate a teacher's *goals* in subsequent post-learning situations rather than to produce a literal repetition of the demonstrator's actions. For example, if the demonstrator replaces a failing disk drive, CERIL must do the same thing, even if the spare drive has to come from a different location, and even if the faulty drive is in a different slot. CERIL may use a different arm for certain steps, or transfer objects from one "hand" to another, even though the demonstrator did not take these specific actions.

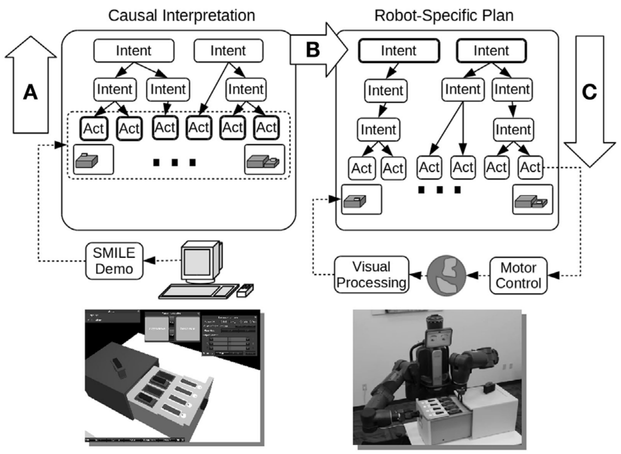


FIGURE 1 | A top-level view of CERIL, the cognitive portion of our imitation learning robotic system. The abductive reasoning processes (infer the causes from the effects) are shown on the left: they produce a hierarchical causal network that represents at the top an *explanation* for the observed demonstrator's actions. After learning, this explanation can be used to guide plan generation in related but modified situations, as illustrated on the right. Figure from Katz et al. (2017a).

As illustrated at the bottom left in **Figure 1**, a person provides a demonstration to CERIL by using a graphical computer program with GUI controls in which the demonstrator manipulates objects on a virtual tabletop (Huang et al., 2015a,b). CERIL uses the event record from this demonstration to infer an explanation for the demonstrator's actions in terms of high-level goals for the shown task (labeled A in **Figure 1**). The high-level goals/intentions/schemas have parameters, such as with *grasp (object*, *location*, *gripper)*. In constructing explanations, CERIL uses predefined goals/intentions and their sub-goals/sub-intentions that are defined *a priori* in its knowledge base. Explanations typically consist of a novel *sequence* of instantiated/grounded high-level goals that CERIL constructs through abductive causal reasoning. In particular, the inference process is an extended version of parsimonious covering theory (Peng and Reggia, 1990). The term "parsimony" refers to the fact that the simplest explanations are to be preferred, while "covering" refers to the fact that a plausible explanation must be able to cause (cover) the observed demonstrator actions. Adapting parsimonious covering as the basis of imitation learning required substantial extensions to the original theory (Katz et al., 2017a). These extensions included incorporating real-valued variables such as object locations and orientations, integrating causal chaining and temporal constraints, and accounting for spatial transformations related to manipulating objects.

Does It Work?

The right side of **Figure 1** illustrates what happens after imitation learning of a task is complete. CERIL can learn and retain multiple tasks over multiple environments, but here we just consider the single disk drive task described above as an example. After learning, CERIL can be given situations in the real world that are similar to what it was trained with (labeled B in **Figure 1**). It will then match its parameterized object models to the objects in the physical environment, which grounds its top-level goals in the new situation. It then uses its grounded explanation (a sequence of goals to be achieved in the order specified) to generate a plan for performing the specific task it has been given by using a hierarchical task network (HTN) planner (Ghallab et al., 2004). This is labeled C in **Figure 1**. From the viewpoint of parsimonious covering theory, this HTN planning process is using CERIL's cause–effect relations in the opposite direction from what was done during learning (i.e., reasoning now goes from causes to effects rather than the opposite which was done during learning). Unlike during the learning phase, HTN planning now involves using goals and actions that are specific to the robot, not to the human demonstrator.

We have systematically tested CERIL using a humanoid physical robot (Baxter, Rethink Robotics™; pictured at the lower right of **Figure 1**) on a set of different tasks, and the detailed results can be found in Katz et al. (2017a). These tasks include learning basic maintenance skills on the disk drive station illustrated above, learning maintenance tasks on a pipe-and-valve plumbing configuration, and learning to construct toy block configurations. In addition, we used computer simulations to test CERIL's ability to interpret correctly action sequences taken from a data set of 5,000 emergency response plans (Blaylock and Allen, 2005). CERIL was able to function effectively and efficiently in all of these situations (Katz et al., 2017a). Most compelling is that CERIL is often able to learn and generalize to modified initial situations (spare disk is in a different initial location, a different indicator light is on, etc.) from a single demonstration, much as a person can do. Further computational simulations comparing different parsimony criteria have investigated the impact of using different criteria for what it is that makes an explanation "parsimonious" (Katz et al., 2017b), and we are currently conducting an experimental study to compare how CERIL's learning and subsequent imitation compare to what is done by human subjects in the same situations.

Finally, a potential benefit of using a cognitive model of the kinds of cause–effect reasoning performed by humans during learning and planning is that it should allow a robot to explain to a human observer why it is carrying out certain actions with justifications that are intuitively plausible. Such an ability is critical to making the simulated reasoning mechanisms of robots and other autonomous systems transparent to people, and this transparency is often an important aspect of machine trustworthiness. We have recently introduced methods by which CERIL can justify its actions to a human observer based on "causal plan graphs" (Katz et al., 2017c). **Figure 2** gives an example of this action sequence justification ability in its current form for a simple device maintenance task. We believe that such "reportability" of underlying inference processes will ultimately prove to be important to investigating the possibility of machine consciousness. The reason for this is that in experimental psychology, investigators long taken a subject's being able to report verbally his/her cognitive experiences to be an objective criterion for that subject to be subjectively aware of those experiences (Baars, 1988; Dehaene and Naccache, 2001).

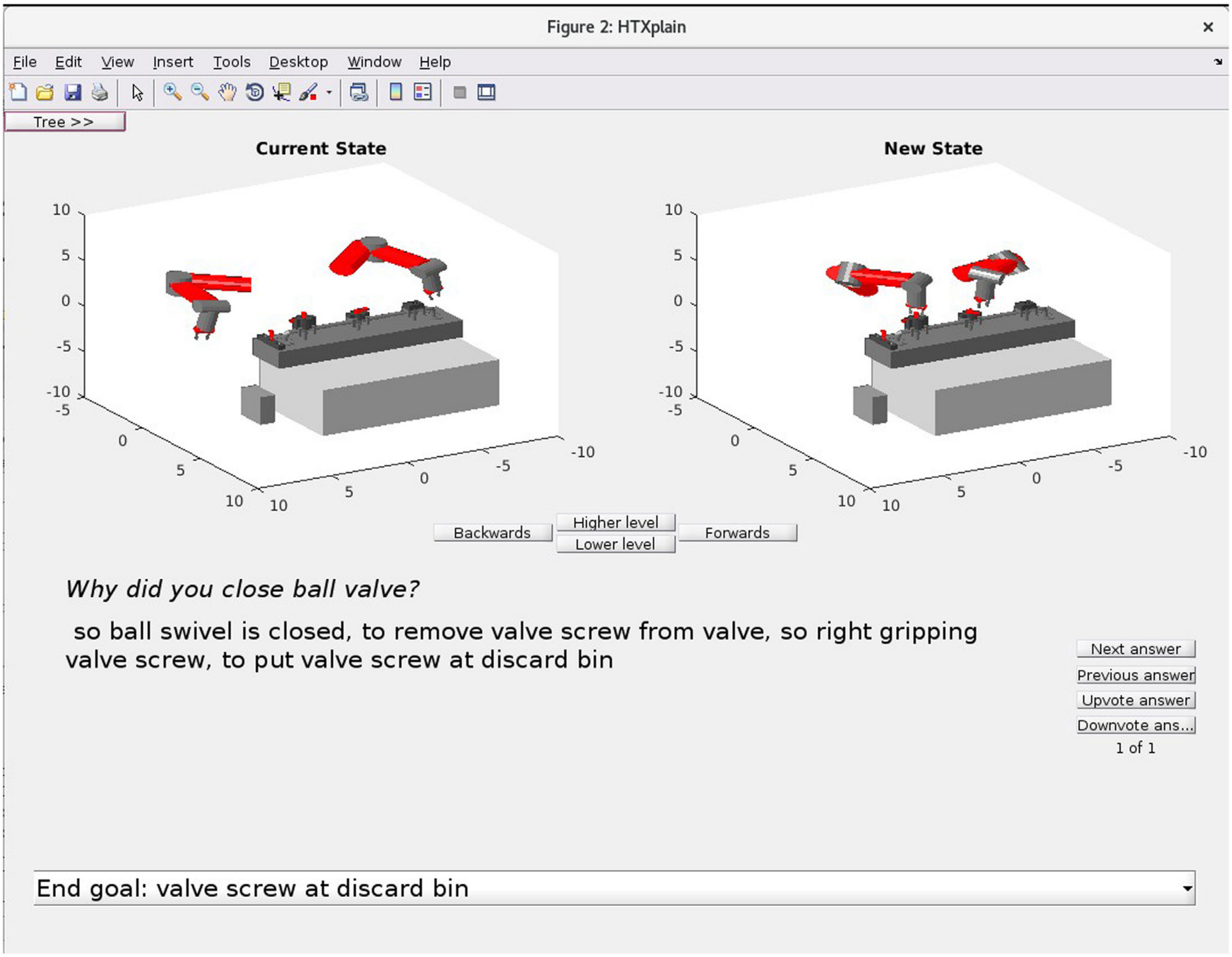
BRIDGING THE CEG

We believe that the imitation learning humanoid robot described above, when controlled by a purely neurocomputational high-level cognitive control system and lower-level sensorimotor system, provides an excellent context in which to study the CEG and to search for potential computational correlates of consciousness. It uses hierarchical causal knowledge, abductive inference, and intention/goal inference processes, all of which have long been widely viewed as modeling important aspects of human reasoning in general and involved in imitation learning specifically (Kassirer and Gorry, 1978; Peng and Reggia, 1990; Josephson and Josephson, 1994; Meltzoff, 1995; Baldwin and Baird, 2001; Bekkering and Prinz, 2002; Haikonen, 2003; Fuster, 2004; Fogassi et al., 2005; Iacoboni et al., 2005; Walton, 2005; Botvinick, 2008; Katz et al., 2017a). However, the control mechanisms instantiated by CERIL are currently implemented with traditional software: Our robot's cognitive components are top-down symbolic AI algorithms for abductive inference and plan generation. In order to use our robotic learning system to study the CEG, the existing software needs to be converted into neurocomputational form, something that is currently in progress. At present, we have converted the low-level sensorimotor control of individual robot actions into neural network modules, replacing the corresponding original software with a neural architecture, the DIRECT algorithm, that we have previously studied *via* non-robotic computer simulations (Gentili et al., 2015). Testing of the resulting robotic control system (i.e., the top-down symbolic cognitive components plus the neural sensorimotor components instantiated in our robot) on tasks such as maintenance operations on the disk drive dock and pipe-and-valve system described above show that the robot's behavior with a neural sensorimotor system is virtually unchanged from the original.

We have concurrently also been studying, so far only *via* non-robotic computer simulations, neural mechanisms for cognitive control of working memory and other behaviors that are intended to serve as purely neurocomputational replacements for CERIL's existing executive control system. In the rest of this section, we first describe the neurocomputational systems we are developing that are inspired by both cortical and subcortical processes that are believed to underpin human cognitive control mechanisms. We then describe a key hypothesis of our work addressing the CEG: that top-down gating of working memory is an important computational correlate of consciousness. This hypothesis is motivated in part by the recognition by many psychologists that working memory is a significant aspect of conscious human cognition, as we explain further below.

Neurocomputational Implementation of Top-Down Gating

The current implementation of our robotic system for imitation learning provides a good illustration of the CEG as we portrayed it above: high-level cause–effect reasoning and planning



Why did you close ball valve?

so ball swivel is closed, to remove valve screw from valve, so right gripping valve screw, to put valve screw at discard bin

End goal: valve screw at discard bin

FIGURE 2 | Because of its use of cause–effect knowledge and abductive inference methods that are arguably models of human knowledge and reasoning, CERIL can generate simple intuitive justifications of its actions to a person who is observing a humanoid robot at work. While the English is a bit stilted, in the example shown here CERIL is responding to a question as to why it closed a ball valve by describing its reasons (causative factors based on its goals).

successfully implemented using symbolic AI operations, and low-level sensorimotor control successfully implemented using neural network methods. Given the framework that we have outlined above (see A Framework for Investigating Consciousness), our specific research agenda is clear: search for computational correlates of consciousness by replacing CERIL's causal reasoning and planning algorithms with a purely neurocomputational system that provides the same functionality. Such a replacement is beyond the reach of current neurocomputational technology and is a very challenging target. However, it provides a concrete example of attempting to bridge the CEG, and in this context it has the potential to reveal candidates for computational correlates of consciousness as per our research framework and cognitive phenomenology.

Given this challenge, we are taking inspiration from what is known about the neurobiological mechanisms underlying human cognitive control. Of course, current understanding of these biological mechanisms is incomplete, but what is known provides a powerful foundation for addressing how CERIL's mechanisms might be implemented using neural computations. Here, we give two examples of the results we have obtained so far using this approach, explaining for each how they relate specifically to the issue of top-down control of cognitive mechanisms.

First, we created and studied a neurocomputational system named GALIS that models executive control functions and can be related to the CEG (Sylvester et al., 2013; Sylvester and Reggia, 2016). We have studied this model in computer simulations, and the goal now is to adapt an extended version of the methods used in GALIS as the top-level neural control mechanisms in CERIL. As illustrated in **Figure 3**, this model is centered on an executive system that gates (turns on or off) the functions of the other components of the system, including working memory. The working memory module is an autoassociative recurrent network that adopts one-step Hebbian synaptic changes to quickly store

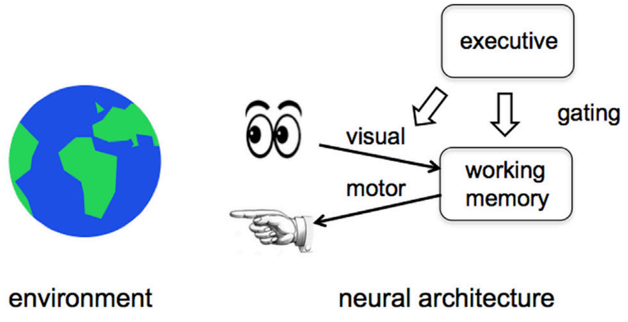


FIGURE 3 | The top-level architecture of GALIS' neural control system. The operational components of an intelligent agent's control system, such as visual information processing, motor control, and working memory, are gated by an executive system at the upper right. Our work focuses on the top-down gated control of working memory in particular.

and recall problem-solving information such as what objects are in the workspace and their locations. The executive control module, of primary interest here, is trained to activate/de-activate the functions of the other components in the system. This gating control mechanism thus determines whether or not inputs are saved in working memory, when information stored in working memory is to be deleted, and when outputs are to be produced. Using Hebbian learning methods, it is possible to "program" GALIS to carry out tasks that require a sequence of motor actions to be executed that are specific to solving a given problem. For example, we trained GALIS to play simple card games that required it to retain in working memory the previous cards that it had seen, and to base decisions about its actions on the contents of working memory. Not only did GALIS perform the task well in solving hundreds of randomly generated card game problems, but it was also found to exhibit some significant similarities to people in terms of how many steps it took to solve card game problems of various difficulty levels (Sylvester and Reggia, 2016) as well as in memory capacity in separate experiments simulating human n-back problem solving (Sylvester et al., 2013).

The executive component, shown at the upper right in **Figure 3**, is the most interesting aspect of GALIS' underlying neurocomputational system in the context of the CEG. It exerts top-down control over the functions of other operational parts of the overall system. This executive has an internal structure that is more complex than illustrated in **Figure 3**. It consists of multiple components, the most important of which is an associative memory that stores task instructions as attractor states. Each instruction indicates which system components should be activated/de-activated (*via* the gating mechanism) at various times during a task in order to solve whatever problem is under consideration. The executive is trained to represent and remember sequences of instructions ("programs") as sequences of attractor states. Like working memory, learning is based on Hebbian synaptic changes. Subsequently, the executive sequentially visits those learned attractor states in the correct order during problem solving. In effect, this procedural memory allows the executive to learn to represent simple tasks (sequences of instructions or "programs") as sequences of transient attractor states. This is of special interest in the context of past suggestions that some activity state trajectories in neural systems might be computational correlates of consciousness (Fekete and Edelman, 2011). What our model adds to this suggestion is the specific idea that temporal sequences of *attractors* (itinerant attractor sequences) used by executive modules instantiating top-down gating might be the specific property that makes activity state trajectories become computational correlates of consciousness. This idea is related to recent work suggesting that sequences of attractor states in recurrent neural networks can shed light on controversies surrounding cognitive phenomenology (Aleksander, 2017). The executive system in GALIS is sufficiently robust even in its current implementation to store and use multiple instruction sequences as appropriate as different conditions arise during problem solving.

Figure 4 elaborates on GALIS' top-level architecture that is illustrated in **Figure 3**. Sensory inputs enter at the upper left, and motor control (e.g., "pointing" at a card) leaves at the bottom left. The internal structure of the recurrently connected networks forming working memory is shown, indicating that this memory stores associated pairs of object-location information. The memory for instruction sequences, or "programs," is a recurrent neural network shown on the right as part of the control module. Not only does it store individual instructions as attractor states (much like the working memory, *via* symmetric synaptic weights produced by one-step Hebbian learning), but it also stores the transitions between one instruction to the next. Representing a sequence of attractor states in memory could be done in various ways, e.g., Tani has suggested that compositionality and discrete action sequences (sequences of a nonlinear neural system's states) can be supported *via* chaotic dynamics (Tani, 2017). In GALIS, sequencing between instructions is instead based on asymmetric weights on recurrent connections in the instruction sequence memory's network. These asymmetric weights are learned *via* temporally asymmetric Hebbian learning. Thus, during performance of a task, the instruction memory goes to an attractor state (an instruction) corresponding to a local minimum of the network's energy function and performs the specific action(s) indicated by that instruction. The underlying energy landscape governing dynamics then shifts, making the current attractor/instruction unstable since it is no longer an energy minimum state. Guided by the learned asymmetric weights, the state of the network then transitions to a new local energy minimum that is the next instruction in the sequence/program. Multiple instruction sequences can be stored simultaneously in GALIS' control memory. The detailed network structure and equations governing GALIS' activity dynamics and synaptic changes during learning can be found in Sylvester and Reggia (2016).

Most importantly for our discussion here, as the executive system transitions through an instruction sequence, it exerts top-down influences on the functionality of other modules in the system. This control is exerted by gating connections leaving the executive system and traveling to other parts of the system. These gating connections originate at the lower right in **Figure 4** and are labeled g_x in the illustration, where g_x is the activity state of connection x. For example, the executive system turns on learning in the working memory, directing working memory to store the currently seen object's identity and location, by having an output

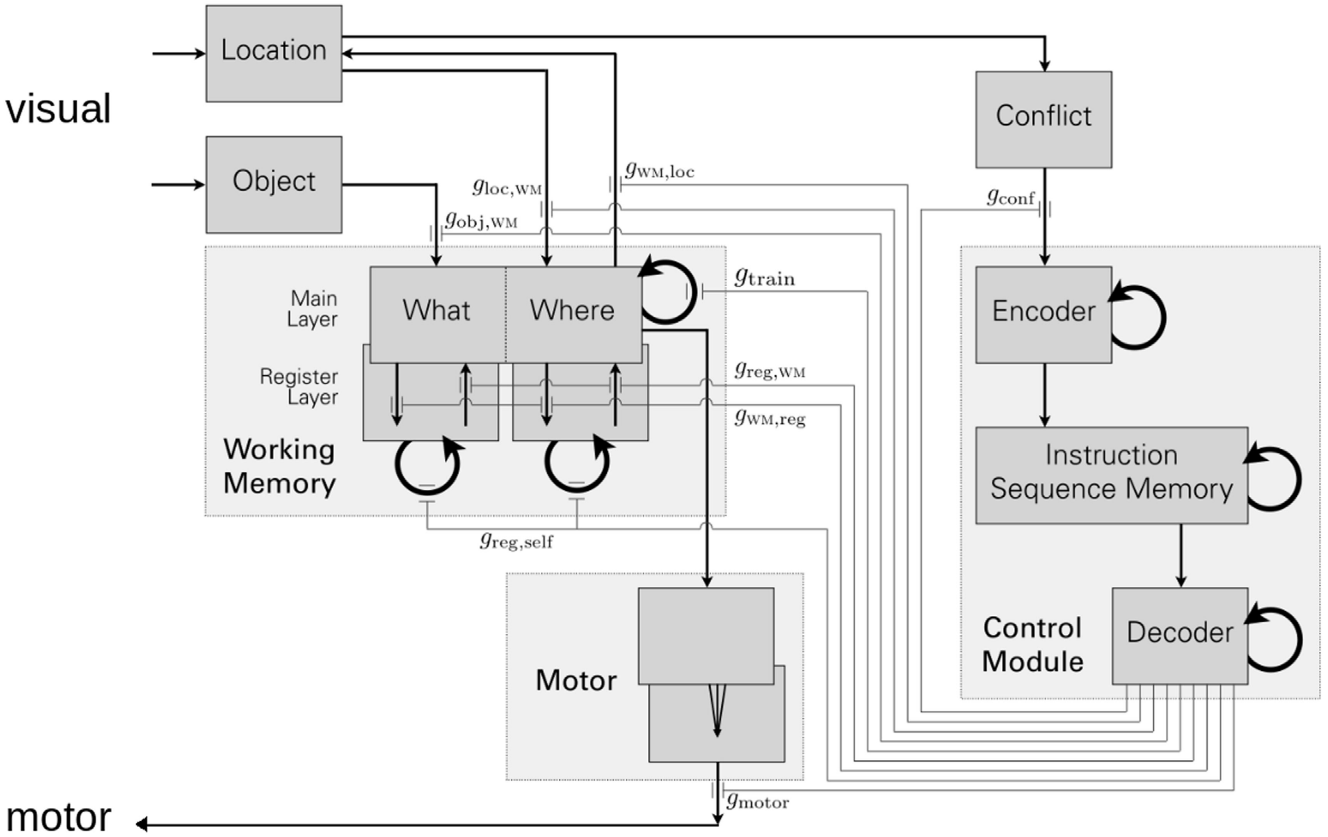


FIGURE 4 | An elaboration of the details of **Figure 3** above. Each rectangle with darker gray shading corresponds to a neural module, where thick semi-circular arrows indicate recurrently connected networks. Thin arrows indicate the connectivity between modules. The working memory of **Figure 3** is depicted on the left, while the executive system (control module) is shown on the right. The connections leaving the decoder module at the lower right of the illustration implement gating actions as explained in the main text. Figure from Sylvester and Reggia (2016).

$g_{\mathrm{train}} = 1$, while it directs working memory to instead ignore the current visual input by having an output of $g_{\mathrm{train}} = 0$. Gating like this in GALIS is implemented *via* "multiplicative modulation" (Akam and Kullmann, 2014), where the g_x values occur in the equations governing activity dynamics and learning in other modules. As an example, if unit k in the motor module has an activity a_k, then what the external world actually sees at that time is the value $g_{\mathrm{motor}} \times a_k$ that incorporates g_{motor} as a multiplying factor. If $g_{\mathrm{motor}} = 1$, then the actual output from unit k at that time is a_k, while if $g_{\mathrm{motor}} = 0$, the actual output is 0. The specific details of how module functionality is gated in the equations controlling system behaviors are given in Sylvester and Reggia (2016).

The core ideas behind GALIS—using top-down gating patterns to encode instructions, and using itinerant attractors to represent sequences of instructions and other data—make for a highly versatile model of computation that can support symbolic reasoning systems like CERIL. For example, suppose that activity patterns are used to represent individual actions and goals that might occur. Itinerant attractor sequences could then be used to store a list of actions that carry out a particular goal, or the list of goals that might cause a particular action, thereby encoding background causal knowledge. Moreover, during reasoning, a working memory could be used to incrementally accumulate a list of conjectured goals that are mutually consistent and account for all actions observed in a demonstration. Finally, instruction memory could be used to store the sequences of gating patterns that carry out the reasoning algorithms. For example, un-gating learning or activation dynamics could be used to store or retrieve background knowledge, respectively. Similarly, during reasoning, un-gated sequence learning in working memory could be used to append new goals when constructing an explanation, and un-gated interactions between background knowledge, working memory, and conflict detection regions could be used to check for inconsistencies before an explanation is modified. Of course, many more subtleties and details will have to be accounted for in a successful implementation. The foregoing examples are intended just to convey the high-level implementation strategy and bolster our claim to its feasibility.

However, a significant limitation to GALIS' executive module is its inability to handle ambiguity. There is no need for a complex decision-making process in the card matching task described above because it could be specified with a simple set of deterministic rules to carry out based on the state of the environment. More realistic tasks often necessitate decision-making to resolve conflicts between potential responses, and often depend on reinforcement learning mechanisms to determine the relative value of these responses. For this reason, our second effort to implement neural mechanisms that could replace CERIL's symbolic algorithms focuses on the role of subcortical structures like the basal ganglia in cognitive control, such as with decision-making and action selection.

Decades of research have implicated the basal ganglia in a wide array of cognitive and motor functions, many of which are associated with conscious processing (Schroll and Hamker, 2013). Most notably, deficits observed in disorders such as Parkinson's disease suggest a role of the basal ganglia in voluntary movement initiation (Wurtz and Hikosaka, 1986), sequential action performance (Benecke et al., 1987; Jin et al., 2014), attention (Tommasi et al., 2015; Peters et al., 2016), and working memory (Lewis et al., 2005; Gruber et al., 2006). In many cases, the functionality of primitive sensorimotor reflexes in Parkinson's disease patients is correlated with increases in cognitive impairment, suggesting a decreased ability to exert top-down control over unconscious behavior (Vreeling et al., 1993). In addition, there is evidence that abnormal inhibition in the striatum of the basal ganglia is associated with the conscious compulsions reported in tic disorders such as Tourette's syndrome (Vinner et al., 2017). Taken together, this evidence suggests that the basal ganglia comprise an important instrument for conscious top-down control over the central nervous system that could offer a number of potential benefits to a neurocognitive system, including a mechanism for biasing attractor landscapes like those used in GALIS toward reward associated trajectories (goal-directed behavior) and controlling the maintenance and capitulation of salient states (working memory). While the neural mechanisms underlying these processes are not fully understood (Goldberg et al., 2013), past computational models incorporating basal ganglia have been shown to capture important behavioral patterns associated with top-down control (Wiecki and Frank, 2013).

We are currently incorporating such a model into GALIS by dividing the executive module into components corresponding to the prefrontal cortex and the basal ganglia. The latter is intended to address the aforementioned ambiguity issues that arise in complex environments by providing a competitive decision-making component that resolves conflicts arising in the former. In addition, such a component functions as a detector of salient states, thereby providing cues for the timing of behavioral execution, serving as a gate on the gating mechanism itself to prevent premature responses or to interrupt ongoing execution when appropriate. This is particularly relevant to the sensorimotor level of imitation learning. As mentioned above, we replaced traditional low-level motion planning with the DIRECT neural algorithm (Bullock et al., 1993; Gentili et al., 2015), which learns in an unsupervised fashion using exploratory "babbling." Much of the robotic motion planning done during imitation learning of maintenance tasks (like those we described above) requires the use of an inverse kinematics solver that determines a joint trajectory for a given end-effector starting position and target. DIRECT learns this coordinate transformation in a self-organizing map architecture by training on a randomly generated set of joint movements and their consequential end-effector transformations. It computes inverse kinematics by finding a difference vector and adjusting the end-effector position using the transformed kinematic information for the appropriate movements that must be made to reach the goal state. Once trained, the resulting model is capable of producing iterative joint movements that approximate the shortest path to the target position.

We have developed an augmented version of the DIRECT model that controls imitation of coordinated bimanual movements (Gentili et al., 2015) to support end-effector orientations, which are critical to performing the demonstrated tasks. This allows the planner to provide joint trajectories that orient the robot's grippers for fine motor tasks, such as manipulating screws, coordinating exchanges of objects between grippers, and fitting objects into tight spaces. However, these additional dimensions were found to pose a unique problem due to the rotational limits of the robot's wrists in the absence of high-level decision-making and top-down control. The DIRECT model is trained to approximate the shortest path to the target position and orientation, but this path may be blocked by the rotational boundary of the wrists, in which case they must be rotated in the opposite direction. Furthermore, a given task may call for a particular rotational direction (for example, unscrewing demands counterclockwise rotation, regardless of the shortest path to the target rotation). Importantly for our work, these considerations motivate the need for top-down control by indicating situations in which top-down control over sensorimotor processing can be used to resolve planning conflicts and override habitual behavior: a gating signal may be used to force the motion planner to take the longer, suboptimal path. It is this kind of context-dependent control over top-down gating that we are currently implementing in the simulated basal ganglia components of our model, which is work in progress.

Top-Down Gating of Working Memory

A key hypothesis of our work addressing the CEG described above is that the top-down gating of working memory (and potentially of other operational components) is an important computational correlate of consciousness. At the least, we believe that studying this aspect of the CEG will lead to the discovery of such correlates. Why is that?

The term *working memory* can be defined as the memory mechanisms that store and process information for a brief period of time. Human working memory has very strict capacity limitations: Psychologists have found that we can only retain about four separate items in our working memory at any point in time (Cowan et al., 2005). If one tries to store more information, the individual items stored may interfere with each other and, in any case, the items will be replaced or decay away over time as problem solving evolves.

The important point here in terms of our work concerning the CEG is that psychologists consider the information processing done by the working memory system to be part of our *conscious* cognitive processes. They have found that storing, manipulating, and recalling information from working memory is conscious and reportable (Block, 2011; Baddeley, 2012). Thus, according to the tenets of cognitive phenomenology (discussed in Section

"Computation, Mind, Brain, and Body"), the computational processes that control working memory deserve consideration as possible computational correlates of consciousness. Further, working memory operations are largely managed *via* cognitive control systems that are biologically most clearly associated with prefrontal cortex "executive functions" that manage other cognitive processes in general (Schneider and Chein, 2003). In terms of the CEG, the issue becomes: can we identify neurocomputational mechanisms that might implement the control of working memory functionality? Elaborating on the hypothesis stated at the beginning of this section, our proposal is that top-down gating like that described above, which determines what is saved and discarded by working memory, furnishes the computational machinery that is used by executive cognitive processes in controlling working memory operations during conscious information processing and is thus a potential computational correlate of consciousness. With top-down gating, an executive module controls the functions of other modules. An executive system may use gating to enable/disable the connectivity between modules, to determine when they remember/forget information, when they generate outputs such as motor actions, and when they learn.

Our specific neurocomputational models described in the preceding subsection envision gating functions, guided by a neurodynamical executive system that sequentially visits attractors that represent instructions (i.e., that represent a procedure for carrying out a task), as corresponding to conscious aspects of cognition that involve working memory. In addition, the gating of working memory in a top-down fashion is reminiscent of the idea of mental causation considered by philosophers deliberating on the topic of free will (Kane, 2005; Murphy et al., 2009). These observations and the finding that control of working memory using top-down gating works effectively in neurocomputational systems and produces behavioral measurement results similar to those observed in humans during n-back memory tasks and card matching tasks (Sylvester et al., 2013; Sylvester and Reggia, 2016) as described above, suggest to us that further investigation of these gating mechanisms may be profitable in the search for computational correlates of consciousness.

DISCUSSION

Current understanding of phenomenal consciousness is widely recognized to be very incomplete, and its relationship to cognition and the core neuroanatomical structures that support it continue to be the focus of recent work (Spreng et al., 2008; Wang and He, 2014; Gomez-Marin and Mainen, 2016). This holds both with respect to consciousness in people and with respect to issues that surround the question of whether machines or animals can be conscious. The primary suggestion in this paper is that the CEG is an important contributing reason for our limited progress toward a better understanding of phenomenal consciousness. This viewpoint runs counter to some past philosophical arguments that understanding the mechanisms of human cognition will not get us any closer to solving the "hard problem" of consciousness. However, the growing recognition among contemporary philosophers who support the idea of cognitive phenomenology suggests, to us at least, that cognition and consciousness are sufficiently intertwined that computational exploration of the CEG may productively lead to insights about the nature of conscious, both in machines and people. It is for this reason that we have suggested a framework for studying consciousness that is based on searching for neurocomputational correlates of consciousness in cognitive-level machines. Ultimately, this general framework, if applied broadly, may turn out to be critically important to providing new knowledge about our basic notions of consciousness. Our view is that the CEG is a central issue for consciousness studies, and one that merits substantial investigation over coming years. Doing this should lead us to discoveries about the computational correlates of consciousness.

More specifically, in this paper we have emphasized the importance of searching for neurocomputational correlates of consciousness, and suggested that one direction in which such a search may prove to be productive is the investigation of executive gating of working memory functions. To our knowledge, very little past work in cognitive robotics or involving computational modeling has examined this specific issue. There have been past computational studies motivated by higher-order thought (HOT) theory that relate cognitive mechanisms to working memory. But these past neurocomputational models based on HOT theory have, to our knowledge, only developed "metacognitive networks" that *monitor* one another, and have not considered the possibility of top-down gating architectures where executive modules *control* other modules' actions. Top-down gating as we describe it here also differs from previously proposed computational models of attention, including proposals that the production of an "efference copy" by control mechanisms (Taylor, 2007) or that having multiple components of a system simultaneously focus on a single subject (Haikonen, 2012), are computational correlates of consciousness. Such models do not explicitly focus on using top-down gating as described in this paper as a control mechanism. As we noted earlier, other past related work includes the suggestion that some activity state trajectories in neural systems might be computational correlates of consciousness (Fekete and Edelman, 2011), and the temporal sequences of attractors used by executive modules instantiating top-down gating in our system is consistent with such a suggestion.

There is much room for further work in this area. For example, at the present time the mechanisms by which a cortical/subcortical region may directly or indirectly control/gate the functions of other regions is not completely clear. Gating interactions in the brain could possibly be implemented by direct pathways between cortical areas, indirectly *via* actions of basal ganglia and thalamic nuclei, by functional mechanisms such as synchronized cortical oscillations, or by some mixture of these and other yet-to-be discovered mechanisms. An important future research topic would be to undertake a more detailed examination of the implications of using alternative gating mechanisms. This relates to the broader issue of what features must be incorporated into computational neural network models to make them adequately representative of brain functions. Current neural network technology spans a broad range of biological realism, running from the relatively realistic Hodgkin–Huxley models incorporating spiking neurons with multi-compartment dendritic trees to the relatively implausible use of linear models or backpropagation learning. In our own work, we have tried to strike a balance regarding this issue, but it remains an important question as to the level of complexity and biological realism in neural computation that will ultimately be best related to the investigation of consciousness. Further future work in neuroscience and psychology is also needed to sharpen our understanding of which cognitive processes are conscious and which are not as a prerequisite for validating computational correlates of consciousness.

AUTHOR CONTRIBUTIONS

All the authors contributed substantially to this work and agree to be accountable for the content of the work. All the authors reviewed and approved the manuscript.

FUNDING

Support for some aspects of this work was provided by ONR award N000141310597.

REFERENCES

Akam, T., and Kullmann, D. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nat. Rev. Neurosci.* 15, 111–123.

Aleksander, I. (2017). "Cognitive phenomenology: a challenge for neuromodelling," in *Proceedings on Aritificial Intelligence and Simulated Behavior*, eds J. Bryson, M. De Vos, and K. Padget (Bath, UK), 395–398.

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

Baars, B. (2002). The conscious access hypothesis. *Trends Cogn. Sci.* 6, 47–52. doi:10.1016/S1364-6613(00)01819-2

Baars, B., Ramsey, T., and Laureys, S. (2003). Brain, conscious experience, and the observing self. *Trends Neurosci.* 26, 671–675. doi:10.1016/j.tins.2003.09.015

Baddeley, A. (2012). Working memory: theories, models and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi:10.1146/annurev-psych-120710-100422

Baldwin, D., and Baird, J. (2001). Discerning intentions in dynamic human action. *Trends Cogn. Sci.* 5, 171–178. doi:10.1016/S1364-6613(00)01615-6

Bayne, T., and Montague, M. (eds) (2011). *Cognitive Phenomenology*. Oxford, UK: Oxford University Press.

Bekkering, H., and Prinz, W. (2002). "Goal representations in imitation learning," in *Imitation in Animals and Artifacts*, eds K. Dautenhahn and C. Nehaniv (MIT Press), 555–572.

Benecke, R., Rothwell, J. C., Dick, J. P., Day, B. L., and Marsden, C. D. (1987). Disturbance of sequential movements in patients with Parkinson's disease. *Brain* 110, 361–379. doi:10.1093/brain/110.2.361

Blaylock, N., and Allen, J. (2005). "Generating artificial corpora for plan recognition," in *User Modeling, LNAI*, Vol. 3538, eds L. Ardissono, P. Brna, and A. Mitrovic (Edinburgh: Springer), 179–188.

Block, N. (1995). On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247. doi:10.1017/S0140525X00038188

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends Cogn. Sci.* 15, 567–575. doi:10.1016/j.tics.2011.11.001

Botvinick, M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* 12, 201–208. doi:10.1016/j.tics.2008.02.009

Bullock, D., Grossberg, S., and Guenther, F. (1993). A self-organizing neural model of motor equivalent reaching and tool use by a multi-joint arm. *J. Cogn. Neurosci.* 5, 408–435. doi:10.1162/jocn.1993.5.4.408

Carberry, S. (2001). Techniques for plan recognition. *User Model. User Adapt. Interact.* 11, 31–48. doi:10.1023/A:1011118925938

Chalmers, D. (1996). *The Conscious Mind*. Oxford, UK: Oxford University Press.

Chalmers, D. (2000). "What is a neural correlate of consciousness?" in *Neural Correlates of Consciousness*, ed. T. Metzinger (Cambridge, MA: MIT Press), 17–39.

Chella, A. (2007). "Towards robot conscious perception," in *Artificial Consciousness*, eds A. Chella and R. Manzotti (Exeter, UK: Imprint Academic), 124–140.

Chella, A., Dindo, H., and Infantino, I. (2006). A cognitive framework for imitation learning. *Rob. Auton. Syst.* 54, 403–408. doi:10.1016/j.robot.2006.01.008

Chella, A., and Gaglio, S. (2012). Synthetic phenomenology and high-dimensional buffer hypothesis. *Int. J. Mach. Conscious.* 4, 353–365. doi:10.1142/S1793843012400203

Chudnoff, E. (2015). *Cognitive Phenomenology*. Routledge Press.

Cleeremans, A. (2005). Computational correlates of consciousness. *Prog. Brain Res.* 150, 81–98. doi:10.1016/S0079-6123(05)50007-4

Cleeremans, A., Timmermans, B., and Pasquali, A. (2007). Consciousness and metarepresentation: a computational sketch. *Neural Netw.* 20, 1032–1039. doi:10.1016/j.neunet.2007.09.011

Cowan, N., Elliott, E., Scott Saults, J., Morey, C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cogn. Psychol.* 51, 42–100. doi:10.1016/j.cogpsych.2004.12.001

Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness. *Cognition* 79, 1–37. doi:10.1016/S0010-0277(00)00123-2

Dindo, H., Chella, A., La Tona, G., Vitali, M., Nivel, E., and Thorisson, K. R. (2011). "Learning problem solving skills from demonstration," in *AGI, LNCS*, Vol. 6830, eds J. Schmidhuber, K. R. Thorisson, and M. Looks (Berlin, Heidelberg: Springer), 194–203.

Fekete, T., and Edelman, S. (2011). Towards a computational theory of experience. *Conscious. Cogn.* 20, 807–827. doi:10.1016/j.concog.2011.02.010

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667. doi:10.1126/science.1106138

Friesen, A., and Rao, R. (2010). "Imitation learning with hierarchical actions," in *Proc. of the 9th Intl. Conf. on Development and Learning* (IEEE), 263–268.

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

Gentili, R., Oh, H., Miller, R., Huang, D., Katz, G., and Reggia, J. (2015). A neural architecture for performing actual and mentally simulated movements during self-intended and observed bimanual arm reaching movements. *Int. J. Soc. Robot.* 7, 371–392. doi:10.1007/s12369-014-0276-5

Ghallab, M., Nau, D., and Traverso, P. (2004). *Automated Planning*. San Francisco, CA: Elsevier.

Goldberg, J. H., Farries, M. A., and Fee, M. S. (2013). Basal ganglia output to the thalamus: still a paradox. *Trends Neurosci.* 36, 695–705. doi:10.1016/j.tins.2013.09.001

Gomez-Marin, A., and Mainen, Z. (2016). Expanding perspectives on cognition in humans, animals and machines. *Curr. Opin. Neurobiol.* 37, 85–91. doi:10.1016/j.conb.2016.01.011

Gruber, A. J., Dayan, P., Gutkin, B. S., and Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *J. Comput. Neurosci.* 20, 153–166. doi:10.1007/s10827-005-5705-x

Haikonen, P. (2003). *The Cognitive Approach to Conscious Machines*. Exeter, UK: Imprint Academic.

Haikonen, P. (2012). *Consciousness and Robot Sentience*. Singapore: World Scientific.

Huang, D., Katz, G., Langsfeld, J. D., Oh, H., Gentili, R. J., and Reggia, J. (2015a). "An object-centric paradigm for robot programming by demonstration," in *Foundations of Augmented Cognition 2015. LNCS*, Vol. 9183, eds D. D. Schmorrow and M. C. Fidopiastis (Springer), 745–756.

Huang, D., Katz, G., Langsfeld, J. D., Gentili, R. J., and Reggia, J. (2015b). "A virtual demonstrator environment for robot imitation learning," in *IEEE Intl. Conf. on Technologies for Practical Robot Applications (TePRA)* (IEEE), 1–6.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3:e79. doi:10.1371/journal.pbio.0030079

Jin, X., Tecuapetla, F., and Costa, R. M. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nat. Neurosci.* 17, 423–430. doi:10.1038/nn.3632

Jorba, M., and Vincente, A. (2014). Cognitive phenomenology, access to contents, and inner speech. *J. Conscious. Stud.* 21, 74–99.

Josephson, J., and Josephson, S. (1994). *Abductive Inference*. Cambridge, UK: Cambridge University Press.

Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford, UK: Oxford University Press.

Kassirer, J., and Gorry, G. (1978). Clinical problem solving: a behavioral analysis. *Ann. Intern. Med.* 89, 245–255. doi:10.7326/0003-4819-89-2-245

Katz, G., Huang, D., Gentili, R., and Reggia, J. (2016). "Imitation learning as cause-effect reasoning," in *9th Conf. on Artificial General Intelligence* (New York, NY: Springer Intl. Publishing).

Katz, G., Huang, G., Hauge, T., Gentili, R., and Reggia, J. (2017a). "A novel parsimonious cause-effect reasoning algorithm for robot imitation and plan recognition," in *IEEE Trans. on Cognitive and Developmental Systems*.

Katz, G., Huang, D., Gentili, R., and Reggia, J. (2017b). "An empirical characterization of parsimonious intention inference for cognitive-level imitation learning," in *Proc. 19th Intl. Conf. on AI* (Los Vegas).

Katz, G., Dullnig, D., Davis, G., Gentili, R., and Reggia, J. (2017c). "Autonomous causally-driven explanation of actions," in *International Symposium on Artificial Intelligence* (Los Vegas).

Kautz, H., and Allen, J. (1986). "Generalized plan recognition," in *Procs. 1986 of the American Association for Artificial Intelligence, AAAI*, 32–37.

Kitamura, T., Tahara, T., and Asami, K. (2000). How can a robot have consciousness? *Adv. Robot.* 14, 263–275. doi:10.1163/156855300741573

Kuipers, B. (2008). Drinking from the Firehose of experience. *Artif. Intell. Med.* 44, 155–170. doi:10.1016/j.artmed.2008.07.010

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pac. Philos. Q.* 64, 354–361. doi:10.1111/j.1468-0114.1983.tb00207.x

Lewis, S. J., Slabosz, A., Robbins, T. W., Barker, R. A., and Owen, A. M. (2005). Dopaminergic basis for deficits in working memory but not attentional set-shifting in Parkinson's disease. *Neuropsychologia* 43, 823–832. doi:10.1016/j.neuropsychologia.2004.10.001

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., and Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science* 309, 2228–2232. doi:10.1126/science.1117256

Meltzoff, M. (1995). Understanding the intentions of others. *Dev. Psychol.* 31, 838–850. doi:10.1037/0012-1649.31.5.838

Murphy, N., Ellis, G., and O'Connor, T. (eds) (2009). *Downward Causation and the Neurobiology of Free Will*. New York, NY: Springer.

Nagataki, S., and Hirose, S. (2007). Phenomenology and the third generation of cognitive science: towards a cognitive phenomenology of the body. *Hum. Stud.* 30, 219–232. doi:10.1007/s10746-007-9060-y

Pasquali, A., Timmermans, B., and Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition* 117, 182–190. doi:10.1016/j.cognition.2010.08.010

Peng, Y., and Reggia, J. (1990). *Abductive Inference Models for Diagnostic Problem-Solving*. New York: Springer-Verlag.

Peters, S. K., Dunlop, K., and Downar, J. (2016). Cortico-striatal-thalamic loop circuits of the salience network: a central pathway in psychiatric disease and treatment. *Front. Syst. Neurosci.* 10:104. doi:10.3389/fnsys.2016.00104

Prinz, J. (2011). The sensory basis of cognitive phenomenology. *Cogn. Phenomenol.* 174. doi:10.1093/acprof:oso/9780199579938.003.0008

Reggia, J. (2013). The rise of machine consciousness. *Neural Netw.* 44, 112–131. doi:10.1016/j.neunet.2013.03.011

Reggia, J., Huang, D., and Katz, G. (2015). Beliefs concerning the nature of consciousness. *J. Conscious. Stud.* 22, 146–171.

Reggia, J., Katz, G., and Huang, D. (2016). What are the computational correlates of consciousness? Proc. 2016 Annual International Conference on *Biol. Inspir. Cogn. Archit.* New York, NY.

Reggia, J., Monner, D., and Sylvester, J. (2014). The computational explanatory gap. *J. Conscious. Stud.* 21, 153–178.

Samsonovich, A., and Nadel, L. (2005). Fundamental principles and mechanisms of the conscious self. *Cortex* 41, 669–689. doi:10.1016/S0010-9452(08)70284-3

Schneider, W., and Chein, J. M. (2003). Controlled and automatic processing: behavior, theory, and biological mechanisms. *Cogn. Sci.* 27, 525–559. doi:10.1207/s15516709cog2703_8

Schroll, H., and Hamker, F. H. (2013). Computational models of basal-ganglia pathway functions: focus on functional neuroanatomy. *Front. Syst. Neurosci.* 7:122. doi:10.3389/fnsys.2013.00122

Searle, J. (2004). *Mind*. Oxford, UK: Oxford University Press.

Spreng, R., Mar, R., and Kim, A. (2008). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode. *J. Cogn. Neurosci.* 21, 489–510. doi:10.1162/jocn.2008.21029

Sun, R. (2002). *Duality of the Mind*. Hillsdale, NJ: Erlbaum.

Sylvester, J., and Reggia, J. (2016). Engineering neural systems for high-level problem solving. *Neural Netw.* 79, 37–52. doi:10.1016/j.neunet.2016.03.006

Sylvester, J., Reggia, J., Weems, S., and Bunting, M. (2013). Controlling working memory with learned instructions. *Neural Netw.* 41, 23–38. doi:10.1016/j.neunet.2013.01.010

Tagliazucchi, E., Chialvo, D. R., Siniatchkin, M., Amico, E., Brichant, J. F., Bonhomme, V., et al. (2016). Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics. *J. R. Soc. Interface* 13, 1–12. doi:10.1098/rsif.2015.1027

Takeno, J. (2013). *Creation of a Conscious Robot*. Singapore: Pan Stanford.

Tani, J. (2017). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. Oxford, UK: Oxford University Press.

Taylor, J. (2007). CODAM: a neural network model of consciousness. *Neural Netw.* 20, 983–992. doi:10.1016/j.neunet.2007.09.005

Tommasi, G., Fiorio, M., Yelnik, J., Krack, P., Sala, F., Schmitt, E., et al. (2015). Disentangling the role of cortico-basal ganglia loops in top-down and bottom-up visual attention: an investigation of attention deficits in Parkinson disease. *J. Cogn. Neurosci.* 27, 1215–1237. doi:10.1162/jocn_a_00770

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi:10.1186/1471-2202-5-42

Vinner, E., Israelashvili, M., and Bar-Gad, I. (2017). Prolonged striatal disinhibition as a chronic animal model of tic disorders. *J. Neurosci. Methods* 292, 20–29. doi:10.1016/j.jneumeth.2017.03.003

Vreeling, F. W., Verhey, F. R., Houx, P. J., and Jolles, J. (1993). Primitive reflexes in Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 56, 1323–1326. doi:10.1136/jnnp.56.12.1323

Walton, D. (2005). *Abductive Reasoning*. Tuscaloosa: University of Alabama Press.

Wang, M., and He, B. (2014). A cross-modal investigation of the neural substrates for ongoing cognition. *Front. Psychol.* 5:945. doi:10.3389/fpsyg.2014.00945

Wiecki, T., and Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychol. Rev.* 120, 329–355. doi:10.1037/a0031542

Wurtz, R. H., and Hikosaka, O. (1986). Role of the basal ganglia in the initiation of saccadic eye movements. *Prog. Brain Res.* 64, 175–190. doi:10.1016/S0079-6123(08)63412-3

Conflict of Interest Statement: 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.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 07 February 2018
doi: 10.3389/frobt.2018.00006

SEAI: Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind

Lorenzo Cominelli[1]**, *Daniele Mazzei*[2] *and Danilo Emilio De Rossi*[1]

[1] E. Piaggio Research Center, Information Engineering Department, University of Pisa, Pisa, Italy, [2] Computer Science Department, University of Pisa, Pisa, Italy

A socially intelligent robot must be capable to extract meaningful information in real time from the social environment and react accordingly with coherent human-like behavior. Moreover, it should be able to internalize this information, to reason on it at a higher level, build its own opinions independently, and then automatically bias the decision-making according to its unique experience. In the last decades, neuroscience research highlighted the link between the evolution of such complex behavior and the evolution of a certain level of consciousness, which cannot leave out of a body that feels emotions as discriminants and prompters. In order to develop cognitive systems for social robotics with greater human-likeliness, we used an "understanding by building" approach to model and implement a well-known theory of mind in the form of an artificial intelligence, and we tested it on a sophisticated robotic platform. The name of the presented system is SEAI (Social Emotional Artificial Intelligence), a cognitive system specifically conceived for social and emotional robots. It is designed as a bio-inspired, highly modular, hybrid system with emotion modeling and high-level reasoning capabilities. It follows the deliberative/reactive paradigm where a knowledge-based expert system is aimed at dealing with the high-level symbolic reasoning, while a more conventional reactive paradigm is deputed to the low-level processing and control. The SEAI system is also enriched by a model that simulates the Damasio's theory of consciousness and the theory of Somatic Markers. After a review of similar bio-inspired cognitive systems, we present the scientific foundations and their computational formalization at the basis of the SEAI framework. Then, a deeper technical description of the architecture is disclosed underlining the numerous parallelisms with the human cognitive system. Finally, the influence of artificial emotions and feelings, and their link with the robot's beliefs and decisions have been tested in a physical humanoid involved in Human–Robot Interaction (HRI).

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Luisa Damiano,
Università degli Studi di Messina, Italy
Tony Belpaeme,
Plymouth University, United Kingdom

****Correspondence:***
Lorenzo Cominelli
lorenzo.cominelli@for.unipi.it

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *21 November 2017*
Accepted: *18 January 2018*
Published: *07 February 2018*

Citation:
Cominelli L, Mazzei D and De Rossi DE (2018) SEAI: Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind. Front. Robot. AI 5:6. doi: 10.3389/frobt.2018.00006

Keywords: cognitive systems, artificial intelligence, artificial consciousness, social robotics, humanoids, somatic markers, rules engine, expert systems

1. INTRODUCTION

> Everyone has a rough idea of what is meant by consciousness, but it is better to avoid a precise definition of consciousness because of the dangers of premature definition. Until the problem is understood much better, any attempt at a formal definition is likely to be either misleading or overly restrictive, or both. (Crick and Clark, 1994)

After many years from these words, consciousness is still a thorny and mysterious subject. In human history, almost every philosopher, religious figure, psychologist, and scientist tried to explain its phenomenology. From Plato and Aristotle to Popper and Searle passing through Descartes and Kant, everyone has attempted to pinpoint the "seat of consciousness." Today, this is considered as a process in the body–brain complex, from which consciousness arises and takes shape in terms of attitudes, beliefs, desires, and behaviors. If despite the huge advances in computer science, neurophysiology, and brain imaging, we do not have yet a clear vision about this topic, it is because scientific approaches are very recent. For a long time, consciousness has been perceived as something that is not tangible, not measurable, and therefore impossible to afford by means of scientific methods. Fortunately, nowadays, it is well-known that this assumption depended on a rigid distinction between mind and body, highly affected by cultural and religious convictions; merely, an anachronistic and occidental belief, inherited by the Cartesian division between *res cogitans*, a thinking substance which does not occupy physical space, and *res extensa*, our material body. This theory is no further pursued because of the numerous neuroscientists who demonstrated the strict dependency between our body, emotions, feelings, thoughts, and decisions. In particular, the neuroscientist Antonio Damasio demonstrated how strongly emotions and body are interconnected (Damasio, 1994). His theories were supported by studies conducted on brain-injured patients, thanks to which he disclosed how emotions and feelings emerge through the perception of our body, and how this process is fundamental for the arise of our consciousness (Damasio, 2000).

Another fundamental author, who made an important contribution to the understanding of consciousness, is the philosopher and cognitive scientist Daniel Dennett, with his seminal works "Consciousness explained" (Dennett, 1991) and "Kinds of minds: Toward an understanding of consciousness" (Dennett, 1996). In the former, he denied the existence of a single central place deputed to consciousness (the *Cartesian theater*), describing the brain as a "bundle of semi-independent agencies." In the latter, he led the reader through a fascinating journey in the evolution of living beings to delineate the development of an intelligent conscious mind. He identified this phenomenon with the emergence of capabilities and means that turned out to be advantageous for the interaction between their possessor and the specific environment in which he lives. Therefore, consciousness is explained as the emergence of a set of inner mental representations, which results in the form of intentionality (previously discussed in Dennett (1989)). Clearly, an agent cannot develop any form of intentionality, beliefs, desires, and hence any kind of consciousness, without an autonomous mechanism, which lets him discriminate the entities that share the same environment.

Our purpose is to use an "understanding by building" approach (Webb, 2001) and to treasure all these theories applying them in the field of Social Robotics. In particular, we believe that the Damasio's three-layered theory of consciousness (Damasio, 2000) is applicable as a cognitive model for artificial intelligence (AI) and that the mechanism of somatic markers (Damasio, 1994) is an adequate mechanism for making an artificial agent able to autonomously interpret the entities of its social environment. When followed as design specifications, these can be the key elements to endow a social robot with the possibility to develop more complex and human-like behavior. Such a novel control architecture, highly human-inspired, would be the beginning of a new social robotics control paradigm.

2. COGNITIVE SYSTEMS IN SOCIAL ROBOTICS

There are different definitions of Social Robot (Dautenhahn and Billard, 1999; Bartneck and Forlizzi, 2004; Breazeal, 2004) but they share fundamental characteristics: all these researchers agree that social robots may have different shapes or functions, but they always have to be able to recognize the presence of humans, engage them in a social interaction, express their own synthetic emotional state, and interpret that of its interlocutors. At the same time, they must be able to communicate in a natural human-like way, which should include also non-verbal language, such as communication by gestures, postures, facial expressions, or any other intuitive way. This definition is still true, but after a few years can be not sufficient anymore. Indeed, in the last decade, there has been a massive increase in the diffusion of social robots, and there have been great advances in the fields in which these robots can be involved. Some of these sectors are personal assistance and support in the house of elderly people (Pineau et al., 2003; Broekens et al., 2009; Sharkey and Sharkey, 2012), robot therapy in the hospitals, e.g., in the treatment of ASD disorder (Werry et al., 2001; Pioggia et al., 2005; Scassellati et al., 2012) and depression (Wada et al., 2005; Alemi et al., 2014), contexts of public service (Chung et al., 2007), and even education (Saerbeck et al., 2010; Causo et al., 2016; Vouloutsi et al., 2016). It is evident that their role is moving further and further away from the traditional role of servants, for assuming more the role of companions in a peer relationship. This leads to the need for enhancing some of their requirements, such as empathic behavior, expressiveness, and believability. According to the classification made by Fong et al., it is possible to distribute social robots in a graduated scale that goes from the minimum level of *socially evocative*, robots that rely on the human tendency to anthropomorphize and capitalize on feelings evoked when humans nurture, care, or feel involved with their "creation," to the highest that is *socially intelligent*, robots that show aspects of human-like social intelligence, based on deep models of human cognition and social competence (Fong et al., 2003). The state-of-the-art of this kind of robots shows great results of social robotics in this direction, but, if we focus on the cognitive system controlling a specific robot, it is always characterized by a specific feature that has been highly developed to the detriment of other functionalities.

Reporting some examples of cognitive systems for social robotics, a well-known case is the one of the cartoon-like robot Kismet (Breazeal and Scassellati, 1999). The underlying architecture of this robot was designed on the base of behavioral models and mechanisms of living creatures, and it is referred by Cynthia Breazeal as "the robot's synthetic nervous system" (SNS). This modular framework was structured to provide Kismet with the ability to express lifelike qualities, perceive human social behaviors, and allow the robot to be socially situated with people. Nonetheless, the system was intrinsically designed to model the social interaction between an infant and its caregiver, that resulted in a very sophisticated realism, believability, and expressiveness of the robot, but it did not allow the agent to develop specific behaviors toward different interlocutors neither to reason about their emotional state (Breazeal, 2003, 2004). This work was extended on Leonardo, another robot, whose cognitive system was focused on the functionalities of "perspective-taking" and "mind-reading" (Berlin et al., 2006). An infant-like humanoid that can be definitely considered an emotional social robot is iCub (Metta et al., 2010). It is used as an open-systems platform for research in neuroscience and cognitive development but its biologically inspired cognitive system is more oriented on learning and evolution of some fundamental human movement capabilities, such as object tracking and grasping, or learning by demonstration (Vernon et al., 2007).

In many cases, we found that different approaches correspond to a different level of complexity. For example, a strategy to improve the quality of a social interaction, and increase the empathy of the interlocutors, is to move away from complex cognitive architectures and rely more on the effects of a good affordance, as in the case of Paro (Kidd et al., 2006). The opposite direction has been taken by other researchers, who developed ambitious systems that are highly biomimetic. These research groups are trying to reproduce the function of brain areas and neural pathways for mimicking human cognitive capabilities, as in the case of the Distributed Adaptive Control (DAC) (Verschure, 2012), which has been used in applications with iCub, Zeno (Vouloutsi et al., 2016), and Nao (Fernando et al., 2014).

On the side of artificial consciousness, there is a recent review of cognitive systems inspired by how consciousness arise in humans made by Chella and Manzotti (2013) and another even more recent publication written by Dehaene et al. (2017). We strongly agree with the first authors when saying that consciousness could be the missing step in the ladder from current artificial agents to human-like agents. In the second work, Dehaene et al. suggest that the word "consciousness" conflates two different types of information processing computations in the brain: the selection of information for global broadcasting (C1), and the self-monitoring of those computations (C2). They argue that, despite their recent success, current machines are still mostly implementing computations that reflect unconscious processing (C0) in the human brain. We share also this latter analysis. Indeed, all the cognitive architectures that we investigated are extremely advanced works, and each of these systems, or machines, fully satisfies the purpose for which has been conceived. Nonetheless, in none of these instances, we have found a real creation of personal preferences acquired and processed through the body and emotions of the agent, which is considered the base for the foundation of a potential artificial consciousness.

We identify the best explanation of this process in the Damasio's theory of mind, and we claim that, as yet, the best formalization of this theory is not implemented in any robotic system, but still remains the formalization done by Bosse et al. (2008), which will be introduced in the following section. On the basis of this observation, we decided to design from scratch a novel cognitive architecture for social robotics, which is intended to be the implementation of the Bosse computational model, in order to stay as close to the Damasio's theory of mind as possible. Then, we will test the resulting system to assess the emergence of some form of artificial consciousness and its repercussions on the social behavior and beliefs of an artificial agent.

3. DAMASIO'S THEORY AND ITS COMPUTATIONAL MODEL

In this section, we will cite several parts from Damasio's books (Damasio, 1994, 2000), especially the same parts on which Bosse et al. (2008) focused their attention and took inspiration for their formalization. The theory of mind of Antonio Damasio, as well as the way he described the emergence of consciousness, can be seen as the construction of a building. This construction starts from the emotions, passing through feelings, to arrive to what he calls "feelings of feelings." These are the structural instruments to create the three different levels of consciousness, i.e., respectively: the *proto-self*, the *core consciousness*, and the *extended consciousness*. These three floors share the same building: the body. This latter must be considered not as the theater in which this process takes place, rather, as a necessary means for the generation of consciousness.

According to the general analysis made by Bosse et al. (2008), Damasio described an *emotion (or internal emotional state) as a (unconscious) neural reaction to a certain stimulus, realized by a complex ensemble of neural activations in the brain*. As the neural activations involved often are preparations for (body) actions, as a consequence of an internal emotional state, the body will be modified into an externally observable emotional state. Next, a *feeling* is described as the (still unconscious) sensing of this body state. Finally, *core consciousness* or *feeling a feeling* is what emerges when the organism detects that its representation of its own body state (the proto-self) has been changed by the occurrence of the stimulus: it becomes (consciously) aware of the feeling.

In Damasio (2000), Damasio described this course of events along five steps:

1. *Engagement of the organism by an inducer of emotion, for instance, a particular object processed visually, resulting in visual representations of the object.*
2. *Signals consequent to the processing of the image of the object activate neural sites that are preset to respond to the particular class of inducer to which the object belongs (emotion-induction sites).*
3. *The emotion-induction sites trigger a number of responses toward the body and toward other brain sites, and unleash the full range of body and brain responses that constitute emotion.*
4. *First-order neural maps in both subcortical and cortical regions represent changes in body state. Feelings emerge.*
5. *The pattern of neural activity at the emotion-induction sites is mapped in second-order neural structures. The proto-self is altered because of these events. The changes in proto-self are also*

mapped in second-order neural structures. An account of the foregoing events, depicting a relationship between the "emotion object" (the activity at the emotion-induction sites) and the proto-self is, thus, organized in second-order structures.

Bosse, Junker, and Treur conceived a model, based on these Damasio's notions to simulate the dynamics of the basic mechanisms taking place in the mind and body of an agent. These dynamics are described as an evolution of *states* over time. States are intended as neurological states formed by neural processes. They used the following forms of abstraction:

- neural states or activation patterns are modeled as single *state properties*;
- large multi-dimensional vectors of such (distributed) state properties are composed to one single composite state property, when appropriate; e.g., `(p1, p2, …)` to `p` and `(S1, S2, …)` to `S`.

To describe the dynamics of these processes, Bosse et al. used an explicit reference to time: *dynamic properties* can be formulated relating a state at one point in time to a state at another point in time. They reported the following example "*at any point in time t_1, if the agent observes rain at t_1, then there exists a point in time t_2 after t_1 such that at point t_2 the agent has internal state property s*" (Bosse et al., 2008). Where *s*, in the example, is viewed as a *sensory representation* of the rain. These dynamic properties are expressed in a temporal language, i.e., the Temporal Trace Language (TTL) (Jonker et al., 2003), in which explicit references are made to time points and traces. A *trace* over a state is a time-indexed sequence of states. For performing experiments, they exploited a simpler temporal language called Language and Environment for Analysis of Dynamics by SimulaTiOn (LEADSTO) (Bosse et al., 2005). In this way, they can specify simulation models in a declarative manner. A basic notation of LEADSTO is $\alpha \rightarrow e, f, g, h, \beta$, meaning: "if state property α hold for a time interval with duration *g*, then after some delay (between *e* and *f*) state property β will hold for a time interval of length *h*" (Herlea et al., 1999).

Relying on this descriptive model, they presented a case in which an agent hears some music, which leads to an emotional state that implies physical responses. The process is described by executable Local dynamic Properties (**LP**) in LEADSTO notation, taking into account internal state property `sr(music)` for activated sensory representation of hearing the music, and a vector `p = (p1, p2, …)` of preparation state properties for the activation of the physical responses, defined as the multidimensional composite state property `S = (S1, S2, …)`. A schema of this process is shown in **Figure 1A**, where the corresponding **LP**s are:

LP0 `music → sensor_state(music)`
LP1 `sensor_state(music) → sr(music)`
LP2 `sr(music) → p`
LP3 `p → S`

What is described until **LP3** is the emotional unconscious reaction to a stimulus (or a combination of stimuli), which becomes apparent in the form of bodily changes. According to Damasio (2000), there is still no sense of self nor feelings at this stage, because "*the sense of self has a pre-conscious biological precedent,*

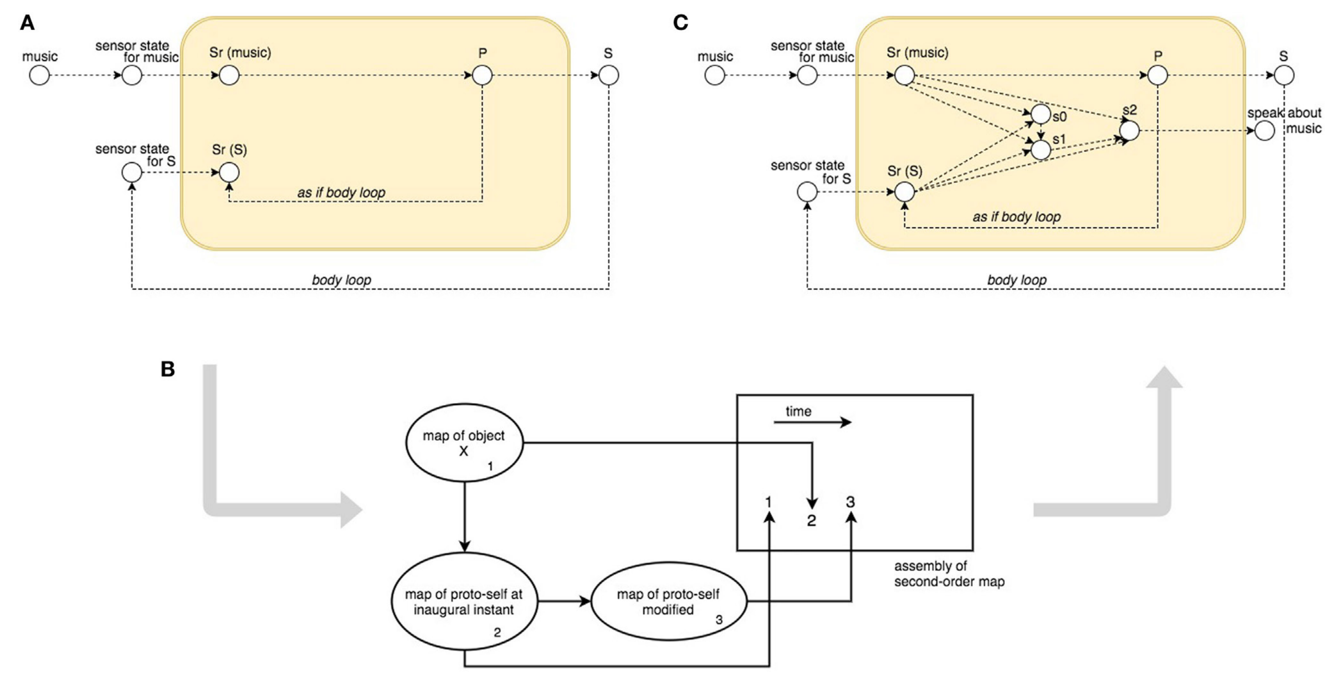


FIGURE 1 | The Bosse et al. computational model: **(A)** *Body loop* and *As If Body Loop* in the generation of feeling; **(B)** Damasio's picture for assembly of a secondary-order map; **(C)** overview of the overall simulation model.

the proto-self, and (…) the earliest and simplest manifestations of self emerge when the mechanism which generates core consciousness operates on that non-conscious precursor."

Here is the point in which body and, particularly, changes in the bodily state perceived as emotions assume their fundamental role for the emergence of feelings, which is described as follows: "*as for the internal state of the organism in which the emotion is taking place, it has available both the emotion as neural object (the activation pattern at the induction sites) and the sensing of the consequences of the activation, a feeling, provided the resulting collection of neural patterns becomes images in mind*" (Damasio, 2000).

Therefore, a feeling emerges when the collection of neural patterns contributing to the emotion lead to mental images. In other words, the organism senses the consequences of the emotional state. This result can be achieved by means of two mechanisms described by Damasio as *via the body loop* and *via the as if body loop*. Bosse, abstracting from the detailed steps made of biological states, summarized these two mechanisms as follows:

Via the *body loop*: the internal emotional state leads to a changed state of the body, which subsequently, after sensing, is represented in somatosensory structures of the central nervous system;
Via the *as if body loop*: the state of the body is not changed. Instead, on the basis of the internal emotional state, a changed representation of the body is created directly in the sensory body maps. Consequently, the organism experiences the same feeling as via the body loop: it is "as if" the body had really been changed but it was not.

This part is formalized including in the model a number of internal state properties for sensory representation of body state properties (`sr(S)`) that are changed due to responses to the stimulus. Together, these sensory representations constitute the feeling induced by the stimulus. As shown in **Figure 1**, `sr(S)` can be reached in two ways, in LEADSTO notation:

LP4 `S → sensor_state(S)`
LP5 `sensor_state(S) → sr(S)`

or

LP6 `p → sr(S)`

where local dynamic properties **LP4** and **LP5** represent the *body loop*, while **LP6** stands for the *as if body loop*.

Finally, Bosse et al. (2008) faced the consciousness problem of "feeling a feeling." Damasio described the origin of consciousness with these words: "*Core consciousness occurs when the brain's representation devices generate an imaged, nonverbal account of how the organism's own state is affected by the organism's processing of an object, and when this process enhances the image of the causative object, thus placing it in a spatial and temporal context (p. 169) (…) beyond the many neural structures in which the causative object and the proto-self changes are separately represented, there is at least one other structure which re-represents both proto-self and object in their temporal relationship and thus represents what is actually happening to the organism: proto-self at the inaugural instant; object coming into sensory representation; changing of inaugural proto-self into proto-self modified by object (p. 177)*" (Damasio, 2000).

Bosse formalized this final part of the process as transitions between the following moments: **(1)** the proto-self at the inaugural instant; **(2)** an object come into sensory representation; **(3)** the proto-self has become modified by the object (see **Figure 1B**). Time is once again the key, and Bosse modeled these steps as a temporal sequence, a *trace*: "(…) in the trace considered subsequently the following events take place: no sensory representations for music and `S` occur, the music is sensed, the sensory representation `sr(music)` is generated, the preparation representation `p` for `S` is generated, `S` occurs, `S` is sensed, the sensory representation `sr(S)` is generated." To model this process, Bosse et al. (2008) introduced three further internal state properties called: `s0` for encoding the initial situation, and `s1` and `s2` for encoding the situation after two relevant changes. The extended model is depicted in **Figure 1C**, formalized by the following LEADSTO notation:

LP7 `not sr(music) & not sr(S) → s0`
LP8 `sr(music) & not sr(S) & s0 → s1`
LP9 `sr(music) & sr(S) & s1 → s2`
LP10 `s2 → speak_about(music)`

The final state `speak_about(music)` is an action made by a conscious agent, who is aware of a feeling, emerged as a change in its body, associated with the specific object that invoked that change. For giving a practical example, thanks to the described process, a person after feeling shivers on his back due to the listening of a song, can make a statement such as the following: *"I love this song,"* where an association has been consciously created between a specific agent (*"I"*), a specific feeling (*"love"*), and a specific evocative object (*"this song"*).

Until this stage of the model, although Bosse states his intention to use a temporal approach, time has not been used. Indeed, the time parameters of LEADSTO (i.e., *e*, *f*, *g*, *h*) are not yet mentioned in the model, which, so far, has a more logical/causal approach. Then, time constraints are reintroduced to allow a simulation of the model. This choice was necessary to allow their software environment to generate traces in the time dimension and, thus, simulate reactions of the model to a controlled sequence of events. They successfully run an experiment in which they simulate both the body loop and the as if body loop. Finally, they deepened the Damasio's concept of "representational content" formalizing in TTL the formation of first-order representations, which refer to external states of world and body, and second-order representations, which refer to internal states (other first-order representations) of the proto-self.

We consider the model proposed by Bosse as the most coherent formalization of Damasio's theory of mind available in the literature. The proof is that we took the mentioned notions as precise instructions for the design of our framework, and numerous references to the model will be made in the next sections. Nonetheless, this model is a purely computational model. It works very well until it is limited to the domain of information processing. When we move to the design of cognitive systems for agents that have to interact in a real environment, new challenging needs and different requirements come out. The real world changes suddenly and unexpectedly, so real-time systems that are involved in real environments must be flexible and always ready to face conflict situations that require solutions. In some cases, the solution has to be quick and responsive. In some other cases, it is required a higher level of reasoning, which can be more abstract, not time-critical, as well as important. In this context, a temporal approach with time constraints is not adequate.

4. THE SEAI FRAMEWORK

> (…) having a mind means that an organism forms neural representations which can become images, be manipulated in a process called thought, and eventually influence behavior by helping predict the future, plan accordingly, and choose the next action. (Damasio, 1994)

The mind is described as a process in which **inputs** from sensors are converted into **knowledge structures** that allow **reasoning**. These inputs can determine immediate **reactions**, while the results of the reasoning process are internal or external **actions** that together with the *newly generated knowledge* drive feelings, emotions, and behaviors of human beings.

Humans perceive the world and their internal state through multiple sensory modalities that in parallel acquire an enormous amount of information creating internal representations of the perceived world. Moreover, behaviors and skills are not innate knowledge but are assimilated by means of a knowledge acquisition process (Brooks et al., 1999) and by emotional influences (Damasio, 1994). This is also supported by the evidence that pure rational reasoning is not sufficient to realize an advantageous decision-making, as demonstrated by studies conducted on subjects with affective and emotional deficits due to brain injuries (Bechara et al., 2000).

SEAI (Social Emotional Artificial Intelligence) is a framework for the development of bio-inspired robotic control systems endowed with a form of artificial consciousness. It is specifically tailored for social robotics applications, where cognitive features aimed at giving agents the capability to perceive, process, and respond to social stimuli are mandatory. Simultaneously, it makes use of the interactions that the agent has with its interlocutors to create beliefs and internal representations that will change its behavior. In order to achieve this purpose, the system has been conceived highly adaptive, responsive but also capable of abstraction and reasoning. As in human nervous system, planning is the slower part of the control architecture. Therefore, the planning engine of the system has been implemented using a rule-based expert system, which can deal with rules and data but is not designed to be fast. In the meanwhile, sensors and actuators deal with quick reactive actions that require fast communication channels and analysis algorithms (Qureshi et al., 2004). For this reason, a hybrid deliberative/reactive architecture, which integrates a rule-based deliberative system with a procedural reactive system, has been selected as main design structure for the SEAI control system.

As shown in **Figure 2**, SEAI services can be conceptually divided into three main functional blocks: **SENSE**, **PLAN**, and **ACT**.

4.1. SENSE

4.1.1. Scene Analyzer

It is the Social Perception System (SPS) that we developed for Social Robots. This service uses dedicated modules that process incoming raw data from sensors (e.g., Microsoft Kinect ONE Camera,[1] TouchMePad (Cominelli et al., 2017), TOI Shield[2]), extract a set of features of the social environment, and contribute to creating integrated "meta-maps," i.e., XML files that include structured information. For example, a *meta-scene* is a structured description of the perceived social environment *(exteroception)*. The extracted features include a wide range of high-level verbal/non-verbal cues of the people presents in the environment, such as facial expressions, gestures, position, age, and gender, and a set of the visually relevant points of the scene calculated from the low-level analysis of the visual saliency map. Finally, the meta-scene is serialized and sent over the network through its corresponding YARP port. Details of the Scene Analyzer algorithms and processes are reported in Zaraki et al. (2017).

4.1.2. Power Supply

It is the energy monitor of the robot. This service manages the connection with the robot power supply and monitors the current consumption and the voltage levels. The Power Supply Monitor (PSM) service calculates the robot power consumption in Watt with a frequency of 1 Hz and serializes this information to be sent over the network. Data coming from PSM constitutes part of the data used to build structured descriptions of the robot's body state *(proprioception)*.

4.2. ACT

4.2.1. Robot Control

This service is the first part of the robot actuation system. Its role is the translation of high-level instructions coming from the deliberative system in low-level instructions for the animators. It has internal modules dedicated to single parts of the robot (e.g., hands, arms, neck, and face). An example of these modules is *HEFES* (Hybrid Engine for Facial Expressions Synthesis), which is a module devoted to emotional control of a facial robot, described in our previous work (Mazzei et al., 2012). This module receives an ECS (Emotional Circumplex Space) point `(v,a)`, expressed in terms of *valence* and *arousal* according to the Russel's theory called "Circumplex Model of

[1] https://developer.microsoft.com/en-us/windows/kinect/hardware.

[2] http://www.besos.cc/portfolio-articoli/toi-shield/.

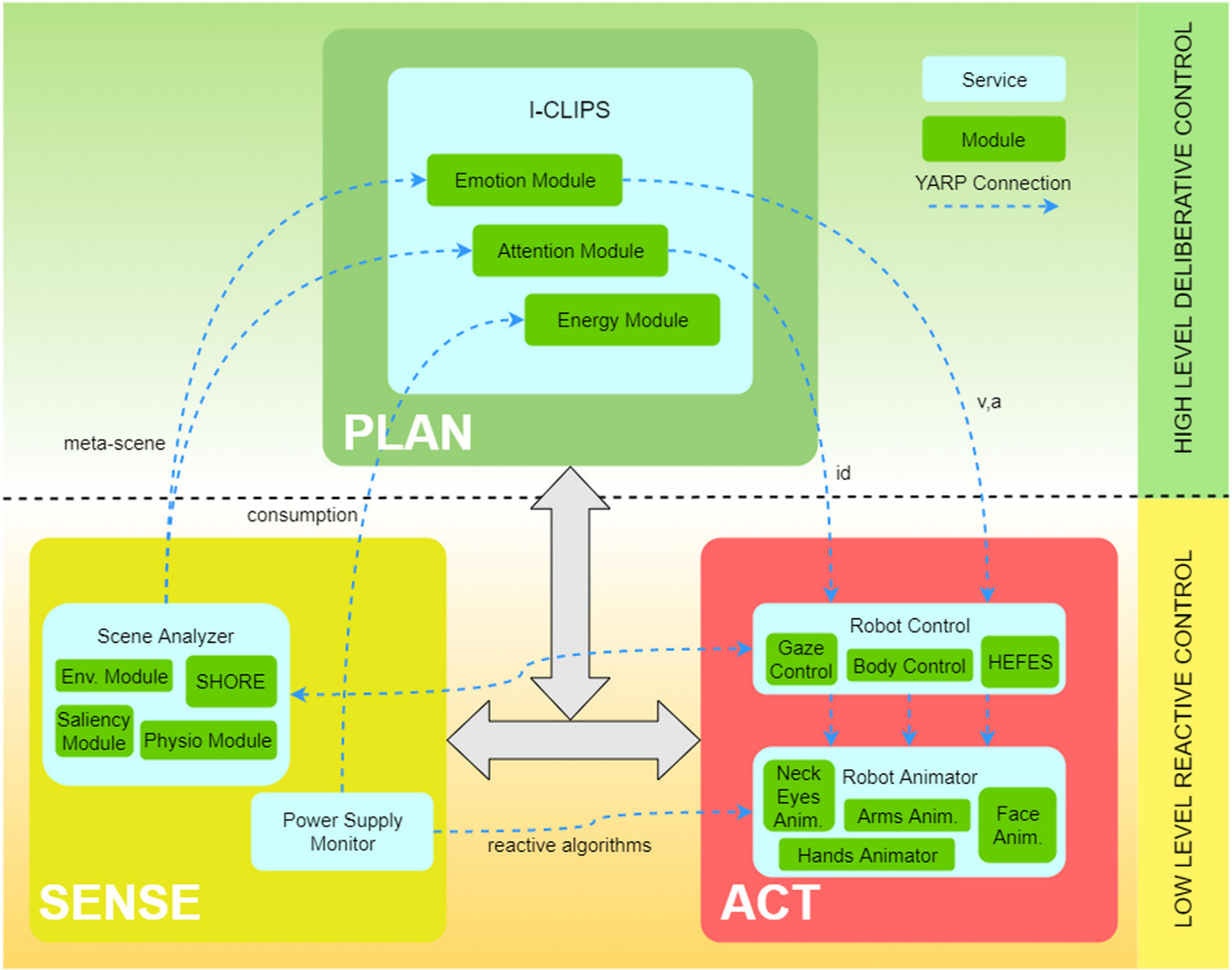


FIGURE 2 | The SEAI architecture includes a set of *services* (blue boxes), standalone applications interconnected through the network. The network communication and services deploy is based on YARP, an open-source middleware designed for the development of distributed robot control systems (Metta et al., 2006). Each service has its *modules* (green boxes) that collect and process data gathered from sensors or directly from the network and send new data over the network. The information flow is defined by XML packets, a serialized form of structured data objects. Thanks to this information management, SEAI is modular and can scale up by developing services, which can even be implemented in different programming languages and placed in different hardware devices. In the proposed architecture ACT, SENSE, and PLAN blocks are only descriptive constructs. The virtual link created by the connections between ACT and SENSE services represents the reactive subsystem. Conversely, the deliberative subsystem is represented by the connections between the I-Clips Rules Engine (PLAN) service and all the other services.

Affects" (Russell, 1980; Posner et al., 2005), and calculates the corresponding facial expression, i.e., a configuration of servo motors that is sent over the network to the Robot Animator. Another example is the module for the *Gaze Control* of the robot, described in details in Zaraki et al. (2014). This module receives directly from the SENSE block a meta-scene object, which contains a list of the persons, each of them identified by a unique `id` and associated with spatial coordinates `(x,y,z)`. The Gaze control module is also listening to the YARP port used by the deliberative subsystem to send the subject's `id` toward which the robot must focus its attention. As a result, the module sends directives to the Neck/Eyes Animator to move the gaze of the robot toward the selected subject.

4.2.2. Robot Animator

It is the low-level service for the actuation of the robot. This service receives multiple requests coming from the *Robot Control*, such as facial expressions and neck movements. Since the behavior of the robot is inherently concurrent, parallel requests could generate conflicts (e.g., a surprised facial expression while blinking). Thus, the Robot Animator is deputed to the distribution of requests through each dedicated animator (e.g., hands animator, face animator, neck/eyes animator, etc.). Moreover, the animation engine is responsible for blending multiple actions taking account of the time and priority of each incoming request. This actuation service is directly connected with the motors moving the robot.

When a service of the ACT block receives an instruction coming from the PLAN block, as the example of an emotion to be expressed, then a deliberative action is taking place. On the contrary, when the instruction is a quick communication due to algorithms that link information gathered by sensors to the movement of motors, the system is dealing with a reactive non-declarative action.

4.3. PLAN

4.3.1. I-CLIPS Brain

The name stands for *Interactive CLIPS*, it is the core of the PLAN block and embeds a rule-based expert system that works as a gateway between the reactive and the deliberative subsystems. The I-CLIPS Rules Engine has been designed using CLIPS (Giarratano and Riley, 1998), and it can be considered as the evolution of our previous work described in Mazzei et al. (2014). In CLIPS expert systems, *facts* represent pieces of information and are the fundamental unit of data used by *rules*. Each fact is recorded in the *fact-list*. I-CLIPS supports the definition of *templates*, structured facts defined as list of named fields called slots. Templates in a declarative language are structured data similar to objects in a procedural language; therefore, it is possible to convert objects in I-CLIPS templates and vice versa. The decision-making process is based on the evaluation of rules. Each rule is composed of two parts: left hand side (*LHS*) contains all the conditions to make the rule trigger, and right hand side (*RHS*) contains the actions that will be fired if the *LHS* conditions are all satisfied. The *RHS* can contain function calls, assertion of new facts or modifications of templates. Assertion of new facts generates new knowledge that can be sent to the other services through the network or used as input for the other rules. If the LHS of a rule is satisfied, that rule is not executed immediately but it is marked as *activated*. Activated rules are arranged in the *agenda*, a list of rules ranked in descending order of firing preference. Rules order in the agenda drives the execution order. Here, the I-CLIPS modules are CLIPS modules (some examples in **Figure 2**). Therefore, each module is a.`clp` file that includes definition of rules and templates. Once a module is loaded by the I-CLIPS Rules Engine, these rules and templates are defined and become part of the SEAI *Knowledge Base*. Modules are distinguished for their function. They have their own agenda and can work in parallel receiving, processing, and sending information through the network. Incoming data can be shared between more modules, as in the case of the *Emotion Module* and the *Attention Module* in **Figure 2**, receiving both the meta-scene, for sending different information in the network, or, no information at all, e.g., the *Energy Module*, because the outcome is a modification of internal parameters (*templates*). The modular structure of the SEAI system allows to include or exclude entire modules, and so, to unable and disable functions at run-time. Modules can have dependencies on other modules, for example, in the rules LHS of module B there can be checks about the state of templates defined by module A. If module A has not been loaded, then module B will not work, but this will not lead to any further consequences. More in general, an activation of an existing function (loading an existing module), or an addition of a new function (loading a new designed module), will not compromise the smooth functioning of the whole system.

What has been described is mainly a causal approach, similar to other approaches in the literature (Manzotti, 2006; Seth, 2008; Chella and Manzotti, 2013), but it is also possible to have partial control on time, in two ways: "prioritization" and "dummy facts." Prioritization of the rules disposition in the agenda can be done declaring *saliency* inside the rules. Saliency is a real number from −10,000 to 10,000 that can be declared in the definition of a rule. Activated rules with higher saliency will be placed at the top of the execution list. No declaration of saliency means saliency equal to 0. With this method, layers of rules inside a module can be created. A layer, which can be considered a sub-module, is a set of rules with the same saliency that connect two or more templates, and it is called a *Rule Set*. In this way, we know that a modification of template T1 will cause a modification of template T2, and not vice versa (if not needed). If multiple rules of the same rule set are activated, they will be ordered on the agenda depending on the selected *conflict resolution strategy*. CLIPS makes available the selection of various conflict resolution strategies among which the *depth strategy* has been selected for its similarity to the typical human reasoning strategy. Using depth strategy, the last rule activated by the facts is the first to be executed generating a behavior that is more responsive and influenced by recent events. The other method is by using "dummy facts." In this latter case, the execution order of rule sets is guaranteed by the assertion of facts: a fact (a *dummy* fact) is asserted as an action of all the RHS of the rules of the precedent rule set and as a condition in the LHS of all the rules of the subsequent rule set, which then will immediately remove that fact from the fact-list, hence the name "dummy."

5. PORTING THE COMPUTATIONAL MODEL IN THE SEAI FRAMEWORK

With respect to the explained framework, we developed modules aimed at replicating the biological mechanisms of consciousness as described by Damasio and then formalized by Bosse. In this section, we present the developed cognitive system dividing the description into the same three notions of "emotion," "feeling," and "feeling of a feeling," and we illustrate how these three levels can be exploited in SEAI for the emergence of the three-layered consciousness defined by Damasio. The "body loop" and the "as if body loop" are also discussed. Moreover, our model of the somatic marker mechanism, which was not included in the Bosse model, will be also described.

First, in order to explain how the SEAI Cognitive System processes the information, another kind of schematic representation is required. Indeed, the functioning of SEAI, akin to the human brain, resides in the structure, meaning the connections among its internal functional parts. In our case, we have a structure made of *templates* connected together by *rules*. The three level of consciousness will be described by gradually loading *modules* that will define templates and rules in the SEAI knowledge base. This schematic representation is highly inspired by the Bosse model (**Figure 1**), where *sensory states* are *templates* or *facts* in our system, and *local dynamic properties* are *rule sets*.

In **Figure 3**, the entire SEAI Cognitive System is shown, where all the developed modules have been loaded.

5.1. The External World

In **Figure 3**, the line delimiting the big white box represents the edge of the physical body of the robot, the gray box in which it is immersed is the external world. Sensors and actuators are the interfaces by which the robot connects with the world. They are represented by a collection of triangles standing in the middle between the body of the robot and the world. Incoming yellow triangles are sensors and outgoing red triangles are actuators. The set of sensors and the perception capabilities depend on the features and the equipment of the robot. As represented in the figure, there are external stimuli that can be perceived by the perception system (bright blue circles), while others (pale blue circles) may not have the corresponding sensory channel in the perception system of the robot. In the case of social robotics, stimuli could be different features of the environment (e.g., temperature, noise level, luminosity, and so on), social cues regarding a unique subject (e.g., gender, facial expression, posture, physio parameters, and so on) or characteristics of an object (e.g., shape, color, dimensions, and so on). Usually, each sensor has a dedicated perception module for the pre-processing of extracted raw data. This is similar to the pre-processing taking place in the human sensory channels. Likewise, the actuation system depends on the motor system of the artificial agent. Typical actuators are servomotors and a set of motors corresponds to a body part of the robot driven by a dedicated animator. However, also speakers for speech synthesis or lights simulating blushing of the skin are considered here as actuators. Arrows coming out from actuators represent the actions of the robot that will lead to some change in the world, this change will be reacquired by the agent as a new collection of external stimuli.

5.2. The Internal World

In the model of **Figure 3**, the focus is all on the PLAN block, which has been extended and its internal structure revealed. The SENSE and ACT block have been compacted in two representational bars with the same reference colors used in **Figure 2**: the yellow bar represents the sum of all perception services, while the red bar stands for the actuation services. Blue boxes are *templates*, and continuous arrows are *rule sets*. Directions of arrows represent the causal/temporal direction due to the abovementioned layering approach. In parallel with external stimuli, the agent has also internal stimuli. They are represented in the schema as an inner blue circle and can be a collection of simulated physiological parameters or a set of values representing the psycho-physical state of the agent. Internal stimuli are updated after every execution cycle after processing the information coming from the external and internal world of the agent. In the middle of the picture, it can be noticed a gray square containing three representative layers. The gray space is the working memory of the robot and corresponds to the "fact-list," the list of all the facts of which the agent is aware of itself and the world. The three representative layers are a symbolic representation through which we describe the arise of consciousness that is reached and enriched by the awareness of facts of increasingly higher level of abstraction. Non-continuous arrows are not rule sets but YARP connections with other services or another kind of connections. These details will be clearer with the following description of rule sets and modules.

5.3. Rule Sets and Modules

Following the key numbers in **Figure 3**: **(0)** external stimuli reach the SENSE block passing through sensors; these connections indicate the sensory acquisition, pre-processing, and integration. These two latter processes take place in the SENSE and provide a single structured meta-map (e.g., a meta-scene)

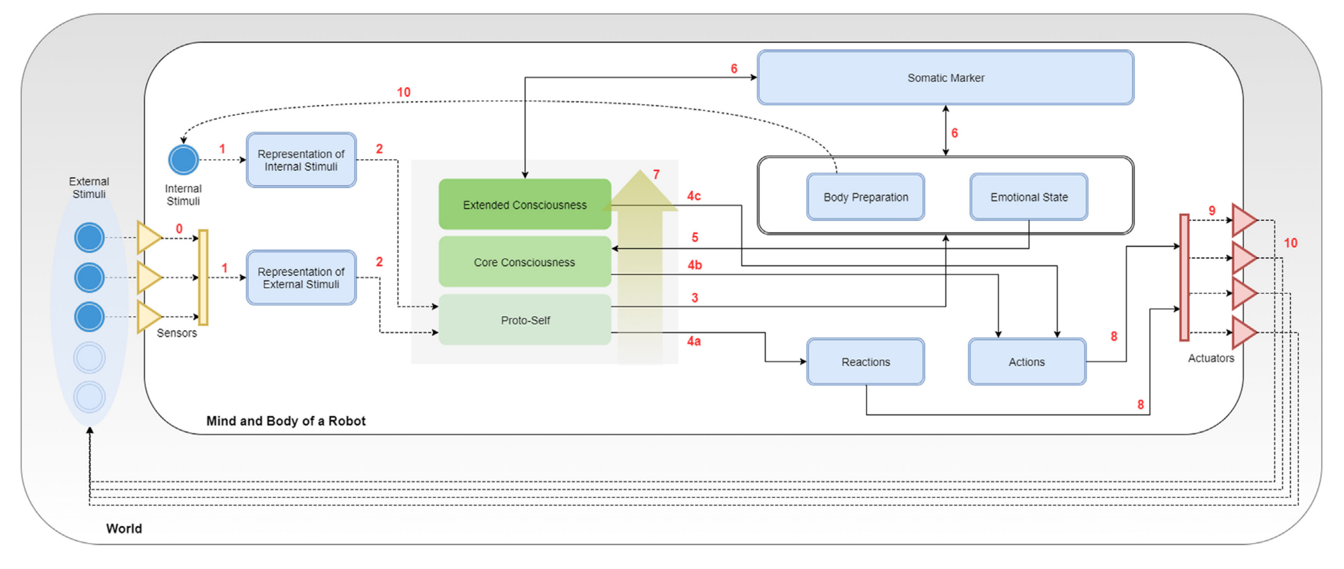


FIGURE 3 | Porting Bosse in SEAI. Key numbers are used for description in section 5.3.

that is sent through a YARP connection. Once the information has been extracted by the external world (exteroception) or perceived from the body (interoception) forming meta-maps, these are analyzed by the deliberative system. **(1)** The system uses pattern matching to compare incoming information with internal representations (pre-defined templates) and recognize real and useful information from inconsistent and useless data. **(2)** If a meta-map has an expected structure and satisfies conditions about internal data, then it is accepted by SEAI as reliable information, and a new fact is asserted in the agent working memory. Facts in the fact-list activate sets of rules of the I-CLIPS rules engine, which will modify other templates or create secondary facts. **(3) EMORS** (EMOtion Rule Set) is a set of rules that analyze facts to process a related emotional predisposition, realized as a modification of values of the templates *body preparation* (`bp(v,a)`), *emotional state* (`es(v,a)`), or both. **(4) BEHRS** (Behavioral Rule Set) is the set of rules that analyze the facts to provide instructions for the robot about certain actions to take, the effect of these rules is the modification of the templates *reactions* or *actions*. This rule set is divided into **(4a) STD-BEHRS** (STandarD Behavioral Rule Set), **(4b) ALT-BEHRS** (ALTernative Behavioral Rule Set), and **(4c) SPEC-BEHRS** (SPECific Behaviors Rule Set), which have increasing priority. This distinction will be clearer in the next section. **(5) FEERS** (FEEling Rule Set) analyze the emotional state template to extract a higher level information that is a conscious feeling, the consequence is the assertion of a secondary fact about the mood of the agent. **(6) SOMARS** (SOmatic MArker Rule Set) is the set of rules simulating the somatic marker mechanism. These rules work in two different directions: they can analyze the body and emotional state to trigger the assertion of a somatic marker, and in case of recognition of a marked entity, they can recall the bodily state that the agent "felt" when that entity was labeled. **(7) REARS** (REAsoning Rule Set) is the set of rules that allows reasoning chain and deductive inferences. These rules do not connect specific templates, because they analyze known facts to assert higher level facts. This rule set is extremely useful to do abstract symbolic reasoning and contributes to the modeling of higher levels of consciousness. Thereby, it is represented by a golden arrow inside the fact-list box. **(8) EXERS** (EXEcution Rule Set) must be the last set of rules to be run. Therefore, they have the lowest saliency values and will be placed at the bottom of the agenda. When all the other rule sets have contributed to the modification of the templates, the actions to take have been decided, the EXERS can send instructions to the ACT Block. This is done through function calls in their RHS that send high-level commands in the YARP network. **(9)** These commands are translated by the Robot Control into motor commands and dispatched by Robot Animator to the actuators of the robot. **(10)** Finally, the bodily state induced by the events is upgraded as a new set of internal stimuli, and the actions of the agent lead to a modification of the social environment that is interpreted as a new set of external stimuli. An execution cycle from 0 to 10 lasts 0.33 ms, which is in line with the physiological time needed for passing from an intention to an action (Libet et al., 1983).

The discussed rule sets and templates are arranged in three different modules:

EMOTION MODULE includes the following: *Representation of Internal Stimuli* template, *Representation of External Stimuli* template, *Reactions* template and *Body Preparation* template. As Rule Sets, the Emotion Module includes EMORS, STD-BEHRS, and a few rules from REARS and EXERS;

FEELING MODULE includes the following: *Emotional State* template, *Actions* template, additional EMORS rules that can modify also (or only) the emotional state, ALT-BEHRS, an extension of REARS, and additional EXERS rules for the execution of actions;

FOF[3] **MODULE** includes the following: *Somatic Marker* template, SOMARS, SPEC-BEHRS, and additional rules of REARS.

As can be noticed, there are entire rule sets that are sole property of a module (e.g., SOMARS) and rules of the same rule set that appear in different modules (e.g., EMORS and REARS). In fact, different modules may include rules with similar function, connecting the same templates, or having the same priority.

5.4. Emotion and Proto-Self

Following the narrative process used in Bosse et al. (2008), we start from a SEAI system in which only the *Emotion module* is loaded (**Figure 4**). Included in the Emotion module, there is the *body preparation* template. As mentioned in the description of the SEAI framework, to model emotion we use the ECS (Emotional Circumplex Space) representation (Russell, 1980). An ECS point is described by two coordinates: *v*, *valence*, the quality of an emotion (i.e., positive or negative), and *a*, *arousal*, which is the activation level of an emotion; *v* and *a* are normalized between 1 and −1. Body preparation is described by a (*v*,*a*) point that is a bodily state, induced by events, that corresponds to a specific emotion. This state will be performed by the agent as an immediate reflex and will last only the duration of the emotional stimulus. Let us assume the same example reported in Bosse et al. (2008), an agent hearing and reacting emotionally to music, and suppose that the SENSE block of SEAI includes a simple software for sound analysis. For example, this software is able to extract the music tempo in terms of beats per minute (bpm) and the sound volume (db). Then, referring to **Figure 4**, this example in SEAI would be the following: **(0)** the music (external stimuli) is acquired by the sensors of the agent (microphones), the audio is processed by the application in the SENSE block, which creates a single structured data: a meta-map containing the perceived characteristic of that music. The meta-map is sent as a YARP bottle in the network; **(1)** the meta-map comes to the I-CLIPS Brain, where is compared with the representation of music, a template `(music (bpm) (volume))`; **(2)** if the information is consistent (e.g., a condition could be $bpm > 0$) then the meta-map becomes a fact in the fact-list, otherwise is rejected; **(7)** REARS may be activated by the `(music)` to do reasoning chain and assert facts, such as `(music-genre-is chill-out)` if $70 < bpm < 120$ or `(volume-is low)` if $db < 45$; **(3)** the appearance of a `(music)` fact activates also the EMORS. For instance, EMORS can trigger specific bodily states in relationship to specific volume ranges. This means a modification of *body*

[3] FOF, Feelings Of Feelings.

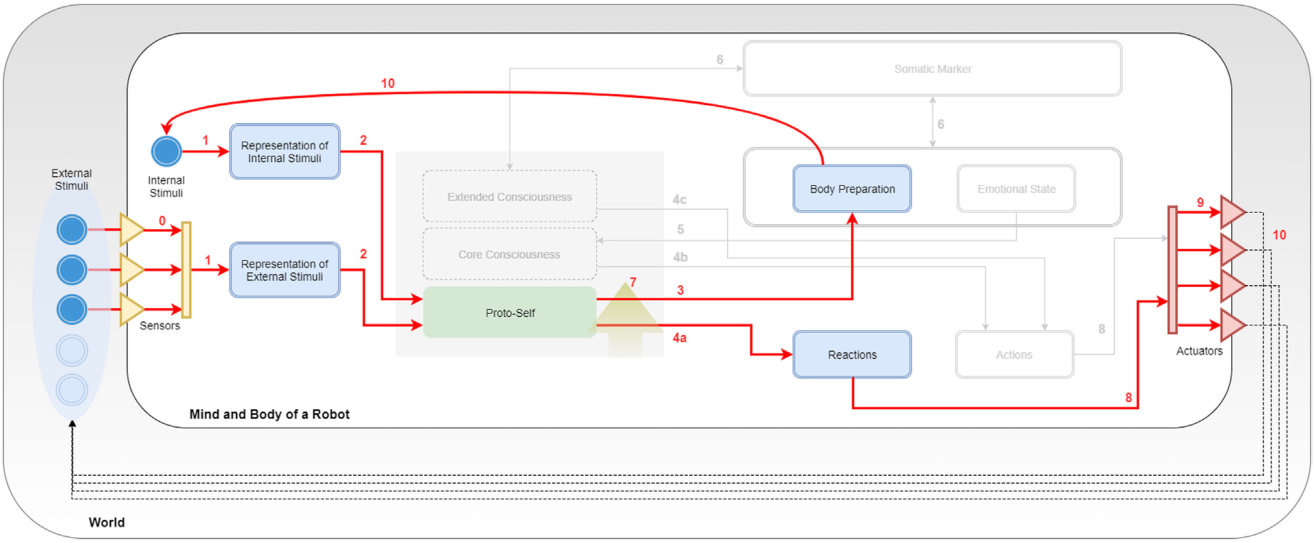


FIGURE 4 | SEAI with only the *Emotion Module* loaded.

preparation from neutral `bp(0,0)` to `bp(v,a)`; **(10,1,2)** this bodily change is updated as an internal stimulus and becomes also a fact in the fact-list; **(4a)** the contemporary presence of the two facts, one about the music and one about the bodily change, activates a behavior, typically a rule of BEHRS which acts on the *reactions* template, copying the bp (v,a) coordinates that now are present as a fact of the fact-list; **(8)** when a disposition is ready and available in the *reaction* template, EXERS is activated and the (v,a) point is sent to the ACT block; services of the ACT block interpret and express the emotional state to perform, translating that emotion in a list of commands for motors. In this way, the emotion is physically expressed through the body of the agent (e.g., a serene facial expression).

This part of the process corresponds to the sequence **LP0**, **LP1**, **LP2**, and **LP3** described in section 3. At this stage, the system is only responsive and capable to process information and express consistent emotional states. The behavior of the agent will be always the same in front of the same stimulus, and its reactions will not last more than the duration of the incoming input. In any case, the simultaneous existence of known facts about the surrounding environment and the body state induced by the entities of that environment fully satisfy the definition of Proto-Self. As a consequence, this first preliminary stage of synthetic consciousness results activated in **Figure 4**.

5.5. Feelings and Core Consciousness

The addition of the *Feeling Module* leads to the definition of new templates and rule sets, which have been highlighted in blue, in **Figure 5**. A new template defined by this module is the *emotional state* template. This new internal representation of the cognitive system is different from *body preparation*. On the one hand, the same emotion model is used for the representation, and so, the instances of this template are also ECS points. On the other hand, `es(v,a)`, unlike `bp(v,a)`, is an internal parameter that does not lead necessarily to an immediate reaction, but rather it is used by the system to modulate the behavior of the robot. This modulation occurs because the module defines new rules of EMORS, which can modify `bp(v,a)`, `es(v,a)`, or both. The `bp(v,a)` points are still discrete states, while `es(v,a)` is modified gradually, by an increase or decrease of its previous (v,a) values. The FEERS checks *emotional state* to assert in the fact-list the current emotional state as a fact. REARS will interpret these states to assert secondary-order facts about the current mood of the agent (e.g., bored, relaxed, and annoyed). The simultaneous presence in the fact-list of a `bp` to perform and an `es` will activate the ALT-BEHRS, which acts on the *actions* template, placing (v',a') values that correspond to

$$v' = (k-1) * v_{bp} + k * v_{es}$$
$$a' = (k-1) * a_{bp} + k * a_{es}$$

where k is the *influence factor*, a global variable, accessible to all modules, which value is set within $0 < k < 1$ and determines the influence of the emotional state on the agent.

Returning to the example of music listening, nothing changes until the sensory representation of the music is asserted as a fact in the fact-list, but now **(3)** new EMORS rules determine variations of the `es` values. For example, there is a rule that makes v_{es} increase together with the music tempo and another one making a_{es} decrease in case of low sound volume. Let us take the case of a slow relaxing music heard at low volume. A protracted listening to this kind of music will lead to: **(5)** the assertion of the fact `es(v,a)` by the FEERS, which every run cycle will be upgraded with decreasing values of both v_{es} and a_{es}; **(4b)** the activation of the ALT-BEHRS due to the contemporary presence of a `bp` and an `es` in the agent working memory; **(7)** the analysis of the `es`-fact by the REARS and the subsequent assertion of secondary-order facts (e.g., `(music-is boring)`). The ALT-BEHRS acts on the *actions* template placing (v',a')

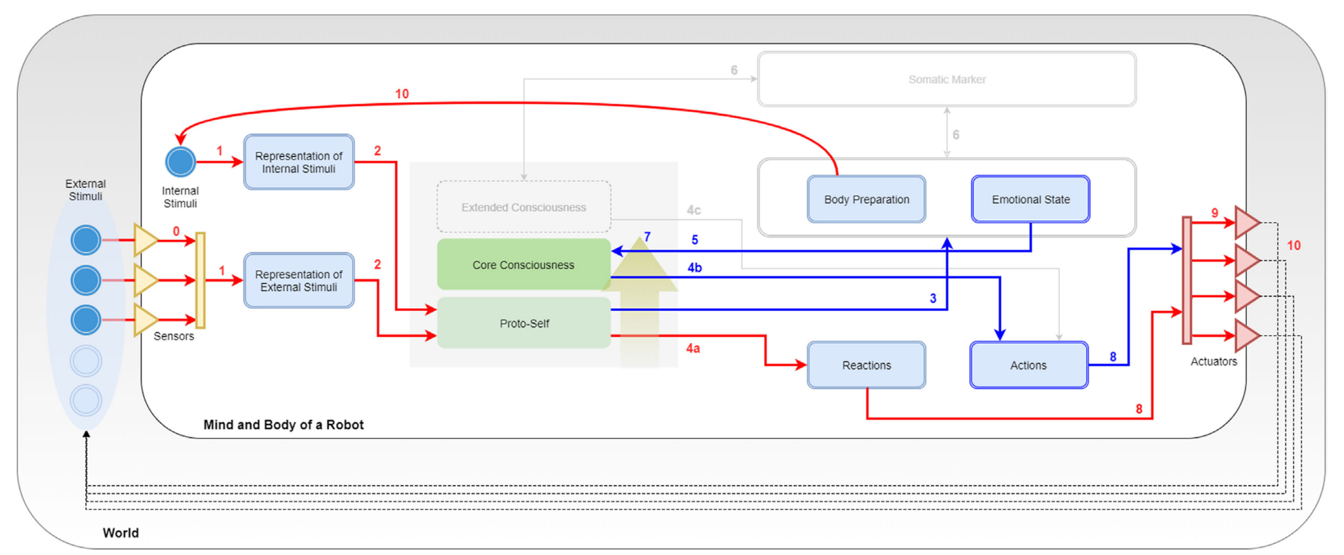


FIGURE 5 | SEAI after *Feeling Module* loading. New parts highlighted in blue.

values. **(8)** The EXERS rules defined by the *Feeling module* have higher saliency than the EXERS rules of the *Emotion module* and check the *actions* template. When all the BEHRS rules have been fired, if both *actions* and *reactions* are filled with values, reactive impulses are temporarily "inhibited" and actions are sent to the ACT block services. The follow-up **(9,10)** is exactly the same described in the previous condition because services of the ACT block are not aware of the declarative process underlying the received instruction. Nonetheless, thanks to *Feeling module*, we will see the previous serene facial expression turning gradually into a bored expression.

The described process corresponds to the addition of **LP4** and **LP5** in the computational model and the emergence from the subcortical to the cortical level in the biological model. It represents the arise of a feeling through the *body loop*. Indeed, the result of this cognitive process is the emergence of secondary-order representations generated by means of slower gradual changes in the body. Here, feelings are not yet internally represented. At this stage, the agent has not a specific behavior toward a precise evocative object, thus, cannot even speak about the music. Nonetheless, reactions to the music are changing, the raised emotions are changing, and feelings are getting clear, which corresponds to the description of what Damasio calls a *Core Consciousness*, that appears activated in **Figure 5**.

5.6. Feeling of a Feeling and Extended Consciousness

In order to uplift feelings and consciousness to a higher level, we relied on the somatic marker hypothesis, formulated by Damasio (1994). A *Somatic marker* (SM) is an association between a relevant change in the body state, perceived as an emotion, and the causative entity that induced that change. According to the hypothesis, somatic markers are processed in the ventromedial prefrontal cortex (VMPFC) and the amygdala and strongly influence subsequent decision-making. Indeed, SMs use our body to create emotional beliefs and opinions about specific entities with which we interact, giving an essential contribute for the formation of an extended consciousness. This mechanism, in case of a second exposure to a marked entity, will recall the body state felt in the past biasing our decisions and behavior toward that specific entity. The hypothesis was demonstrated by Bechara et al. submitting healthy patients and brain-injured patients to the "Iowa Gambling Task," a gambling card game specifically conceived by the authors to assess the efficiency of the SM mechanism (Bechara et al., 1997). To model this brain–body mechanism, we designed the SOMARS. This part of our cognitive system has been tested in a preliminary computational experiment, where we submitted a simulated reproduction of the Iowa gambling task to an artificial agent endowed with SOMARS (Cominelli et al., 2015).

In **Figure 6**, the SEAI system after the loading of the FOF module is shown. This leads to the definition of the *Somatic Marker* template, additional rules in REARS, the SPEC-BEHRS, and SOMARS. SOMARS has been divided into SOMARS rules for SM creation (6a, blue arrows in **Figure 6**) and for SM recall (6b, green arrows in **Figure 6**). To better explain the labeling and recall method, we refer again to the music example: nothing changes in the perception of the music **(0,1)** and the creation of its internal representation as a fact **(1)**; neither the influence of the music on body preparation and emotional state through the EMORS is changed **(3)**, nor the subsequent feelings assertion due to the FEERS **(5)**; but now there are rules of SOMARS that, **(6a)** if the intensity of the emotional state $|es|$, intended as the modulus of $es(v,a)$ vector, exceeds a decided threshold called *sensitivity* (s), then assert a fact in the fact-list: an instance of the *somatic marker* template. A somatic marker in SEAI is a fact `(sm(id) (value)(bp))`, where *id* is an identification number assigned to the causative entity, $value = v_{es} * 100$, and `bp` is a multifield slot that contains the current (v_{bp}, a_{bp}). In the example, the listened

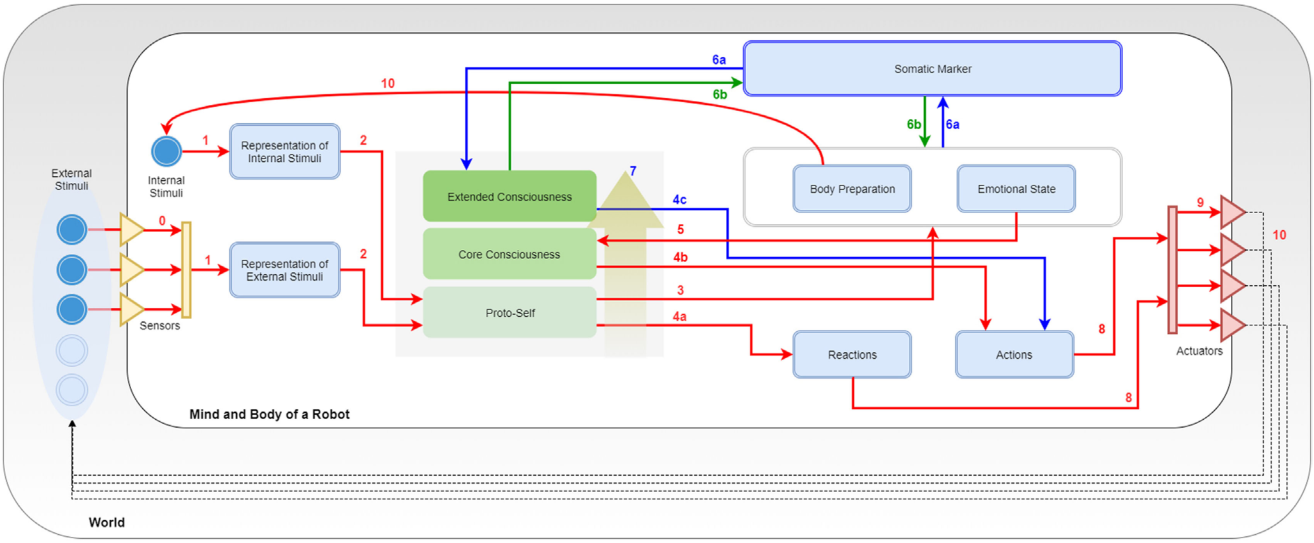


FIGURE 6 | SEAI after *FOF Module* loading. New parts highlighted in blue. Green arrows (6b) indicate SOMARS rules for somatic marker recall.

music, after a few minutes playing, induces by means of EMORS an `es`, which modulus is

$$|es| = \sqrt{v_{es}^2 + a_{es}^2} > s,$$

as a consequence, SOMARS checks the fact-list, the music-genre `chill-out` is identified with a specific `id`, labeled with a `value` and associated with the `bp(v,a)` felt in that moment. A new `(sm)` has been created.

This sequence corresponds to the sequence of transitions between the states *s0* (*the proto-self exists at the inaugural instant*), *s1* (*an object come into sensory representation*), and *s2* (*the proto-self has become modified by the object*). In LEADSTO formalization, this is equivalent to **LP7**, **LP8**, and **LP9**.

From here on, the labeled entity in the fact-list will activate rules of the SOMARS for SM recall **(6b)** that will modify the body preparation state immediately recalling the `bp(v,a)` that was felt and associated with that entity. This `bp` will be represented as a sensory representation of the body state (`sr(S)` in Bosse, a fact in SEAI). This new state is not derived by an upgrade of the body state (**LP4** in Bosse, **10** in SEAI), but from an internal representation of body preparation recalled from the long-term memory of the agent. This is, in all respects, an *as if body loop*, and corresponds in LEADSTO notation to **LP6**.

Another consequence of the recognition of a marked entity may be the activation of **(4c)** a rule of SPEC-BEHRS, triggering some specific behavior toward that entity, pushing a high priority action to be executed, such as saying something about that music (e.g., *"this music is getting boring"*). The sequence that includes **(4c)**, **(8)**, and **(9)** coincides to **LP10**.

Finally, even REARS rules may be activated to assert more abstract and general facts. For instance, a rule of the reasoning rule set could be: if there are the facts `(music)`, `(music-genre is chill-out)`, and a `(sm)` which label that music with a `bp` corresponding to a bored face, then assert the fact `(chill-out is boring)`.

The emergence of SMs is the emergence of personal opinions, about the entities of the world, that the agent autonomously builds through the interactions with such entities. This mechanism, which leads to the construction of an autobiographical memory and biases the behavior of the agent and its opinion about the world, is deputed to the bio-inspired mechanism activated by the FOF module. Things would have ended differently, for example, if other entities of the external world had moved the emotional state in a different direction, predisposing the agent in a better "mood." In this case, chill-out music would have been probably labeled as a nice music genre recalling a pleasant body state to express. In general, it is evident that this level of consciousness, which could not exists without its predecessors, moves beyond the "here and now," includes personal opinions and feelings about specific entities of the world and allows the creation of higher general thoughts. We identify this level with the equivalent of the *Extended Consciousness*, which as a consequence appears activated in **Figure 6**.

6. TESTING SEAI IN THE REAL WORLD—THE HRI EXPERIMENT

In this section, we report an experiment in which SEAI has been used as cognitive system of the humanoid robot FACE (Facial Automaton for Conveying Emotions)[4] (**Figure 7**). FACE is a human-like robotic head, with the appearance of an adult female, capable to perform very sophisticated expressions by means of a hyper-realistic facial mask. The android's head has been customized by our research team starting from a Hanson Robotics[5] head. The facial mask is made of Frubber ("flesh rubber"), a proprietary skin that mimics real human musculature and skin, and

[4] www.faceteam.it.

[5] http://www.hansonrobotics.com/.

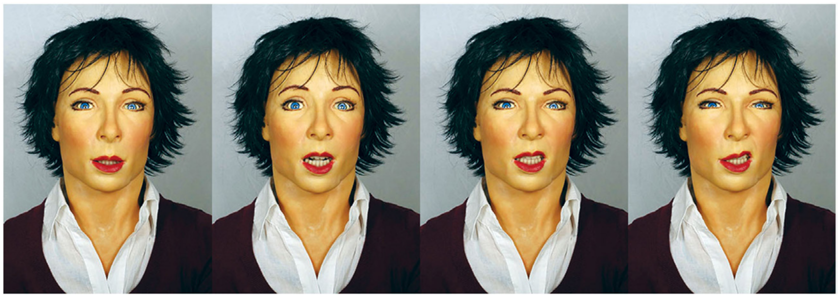


FIGURE 7 | The FACE Robot (Facial Automaton for Conveying Emotions) displaying some of its hyper-realistic facial expressions.

it is actuated by 32 servomotors. The robot has also a mechanical system, composed of a controlled neck with 3° of freedom and movable eyes to allow gaze control (Zaraki et al., 2014, 2017). In this experimental setup, the head has been mounted on a passive mannequin, placed in a seated position. In order to achieve the maximum possible naturalness of the HRI, the interaction takes place in a normal situation of everyday workplace: an office room that has not been prepared or specifically structured. The experiment of this study has been approved by the Ethics Committee of the University of Pisa (prot. 68459, ref. Ethical Approval by CEAVNO, Comitato Etico di Area Vasta Nord). All research participants provided written and informed consent.

In the presented experiment, FACE interacted with three subjects, identified as ID1, ID2, and ID3. The experiment can be divided into the following four scenes:

Scene 1. ID1 enters the room where the robot is seated. He performs several disturbing or impolite actions: he does not greet the robot, immediately invades the robot's intimate space, does not speak to it, folds his arms for a while, and then leaves.
Scene 2. ID2 enters the room and performs mixed actions: he greets robot, invades the robot's intimate space but then immediately makes a step back, speaks for a while to the robot, and then leaves.
Scene 3. ID3 enters the room and performs actions that are typical of nice behavior: he greets warmly the robot, smiles at it, speaks a lot to it; finally, greets again and leaves.
Scene 4. ID1, ID2, and ID3 come back into the room where the robot is located and arrange themselves in three positions at different distance from the robot. They just maintain their position for about 30 s without doing anything to draw the attention of the robot. Then, they all leave the scene.

This sequence has been recorded as a repeatable scenario using Kinect Studio, a tool to record and play back depth, color streams, and audio from a Kinect.[6] In this way, it is possible to present exactly the same scenario to the robot comparing the effect of the same social scene in three different conditions of the cognitive system: (**cond1**) SEAI with only the Emotion module and the *Attention module*; (**cond2**) including the *Feeling module*; and (**cond3**) including the *FOF module*.

Images gathered by the Kinect are analyzed by the Scene Analyzer, which extracts (or estimate) several main social cues of the subjects involved in the scene, e.g., their facial expression, age, gender, gestures, body postures, and proximity. The SENSE service detects also, for every incoming frame, the *salient point* of the image, processed by means of pure image analysis based on colors, contours, light contrast, rapid movements, etc. This point is also identified by an ID, which is ID0. All the information is organized as a *meta-scene* that is sent to the I-CLIPS Brain through YARP. Once the meta-scene has been processed by the I-CLIPS Brain, an ID will draw the attention of the robot that will look at it. This ID is also called *Winner ID*. This is an automatic non-emotional mechanism decided by the rules of the *Attention module*, loaded in all the three conditions. This module, indeed, defines several standard behavioral rules (STD-BEHRS) that, choosing the winner, drive the attention of the robot. For example, the FACE attention is attracted by someone raising their hand or speaking to the robot. If no one is doing anything relevant but subjects are present in the scene, then the robot will look to the closest subject. If no subject is present in the FOV, then the robot will analyze the scene by looking at the salient point. The attention model, here implemented in the form of rules, was studied and discussed in Zaraki et al. (2017).

6.1. Results

6.1.1. Experiment 1

In this first condition, the *Emotion module* is loaded. This leads to the definition of *body preparation* and the EMORS that can modify `bp` (v, a) according to external and internal stimuli. It results in a FACE bodily change, and so, an emotional response to what is happening in its social environment. For example, the absence of people in the FOV of FACE causes the display of a sad facial expression corresponding to negative valence and low arousal (−0.3, −0.5). As the subject enters in the room, we see in **Figure 8** two parallel consequences: rules of the *Attention module* will bias the attention of the robot from the salient point to the detected subject, while rules of the *Emotion module* change the bodily state of the robot. This change in the status of the body will be expressed according to our emotion model through the FACE expressive capabilities: an ECS point is translated by the Robot Control in 32 commands for the relative servomotors moving its face and neck.

[6] https://msdn.microsoft.com/en-us/library/hh855389.aspx.

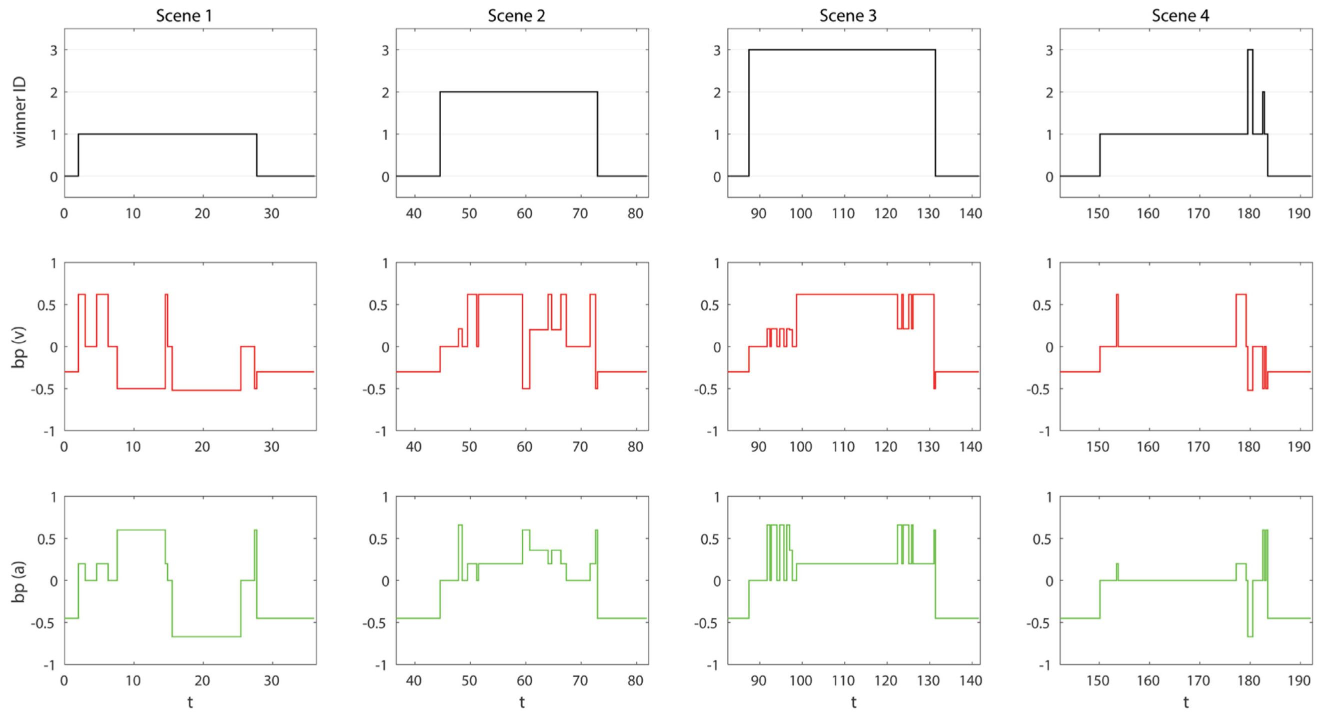


FIGURE 8 | Results of the HRI Experiment with FACE integrating SEAI in condition 1. Columns are the four scenes. Rows are, in order: *winner ID*, *bp (v)*, and *bp (a)*. Time *t* is expressed in seconds.

In **Figure 8**, looking at the charts of `bp(v)` and `bp(a)`, it is possible to see, along all the interaction, the emotional response of the robot. FACE expresses discomfort (−0.5, −0.6) when a subject invades its intimate space, an angry expression (−0.52, −0.67) if someone folds his arms, smiles (0.21, 0.6) if someone greets her or smiles at her, and expresses interest (0.62, 0.2) when an interlocutor speaks to her. Without going into the details of the actions performed by the subjects in their interaction with the robot, the trend of `bp` (*v*,*a*) shows how the robot is emotionally affected in the three first scenes. In the first one, the impolite behavior of ID1 induces unpleasantness and annoyance, hence, values of negative valence are predominant, accompanied by large arousal fluctuations. ID2 has an engaging interaction with the robot, he manifests a polite behavior, quite neutral. As a consequence, positive values of valence are predominant and the arousal is not highly affected. In scene 3, we can see the effects of the interaction with ID3: the interaction is full of positive stimulus, this induce in the robot frequent emotions of pleasantness and high excitement. Finally, we see in scene 4 that, the entire time the robot is detecting people, bodily changes are nearly irrelevant. Indeed, the three subjects just stand in front of the robot without saying or doing anything. The emotion expressed by the robot is always neutral (0,0), with an exception when the subjects leave the room. In this transition, there are fluctuations due to the overlapping of detected people going out through the same door, resulting in a difficult reconstruction of the skeletons by the Scene Analyzer. In any case, sudden quick variations are filtered by the Robot Animator and will not lead to the movement of the robot.

Concerning the behavior of the robot, in terms of attentive model, for the first three scenes, the winners of FACE's attention can only be the single subject presents in each scene or the salient point (ID0). The salient point draws the attention of the robot in the absence of social stimuli, therefore, before and after subjects' detection. In the last scene, including all subjects, the robot focuses its attention on ID1, because he is the closest subject and nobody is doing anything to draw the attention of the robot.

At this stage, FACE bodily state is clearly affected by external events, but the agent is not aware of its own feelings. Emotions last exactly the duration of the stimuli. There is no memory of the experiences. Therefore, behavior is reactive and FACE does not take deliberative decisions about specific subjects. The evidence is that when the subjects come back into the room it is like nothing has happened before, the attention of the robot is not influenced and the robot simply look at the nearest person. We are still at an equivalent of the proto-self level of consciousness.

6.1.2. Experiment 2

The `emotional state` template comes along with the loading of the *Feeling module*. The effects of this module are shown in **Figure 9**. EMORS can now modulate the emotional state (`es`) of the agent, which is continuously upgraded by FEERS through the assertion of facts in the working memory. The influence of events on *es* can vary from a low influence (e.g., talking to the robot, as in scene 3, from $t = 100$ s to $t = 130$ s) to a very important influence (e.g., invading its intimate space, as in scene 1, from $t = 10$ s to $t = 15$ s). This leads to a modification of the emotional state expressed by the robot: the agent does not show exactly the (v_{bp},

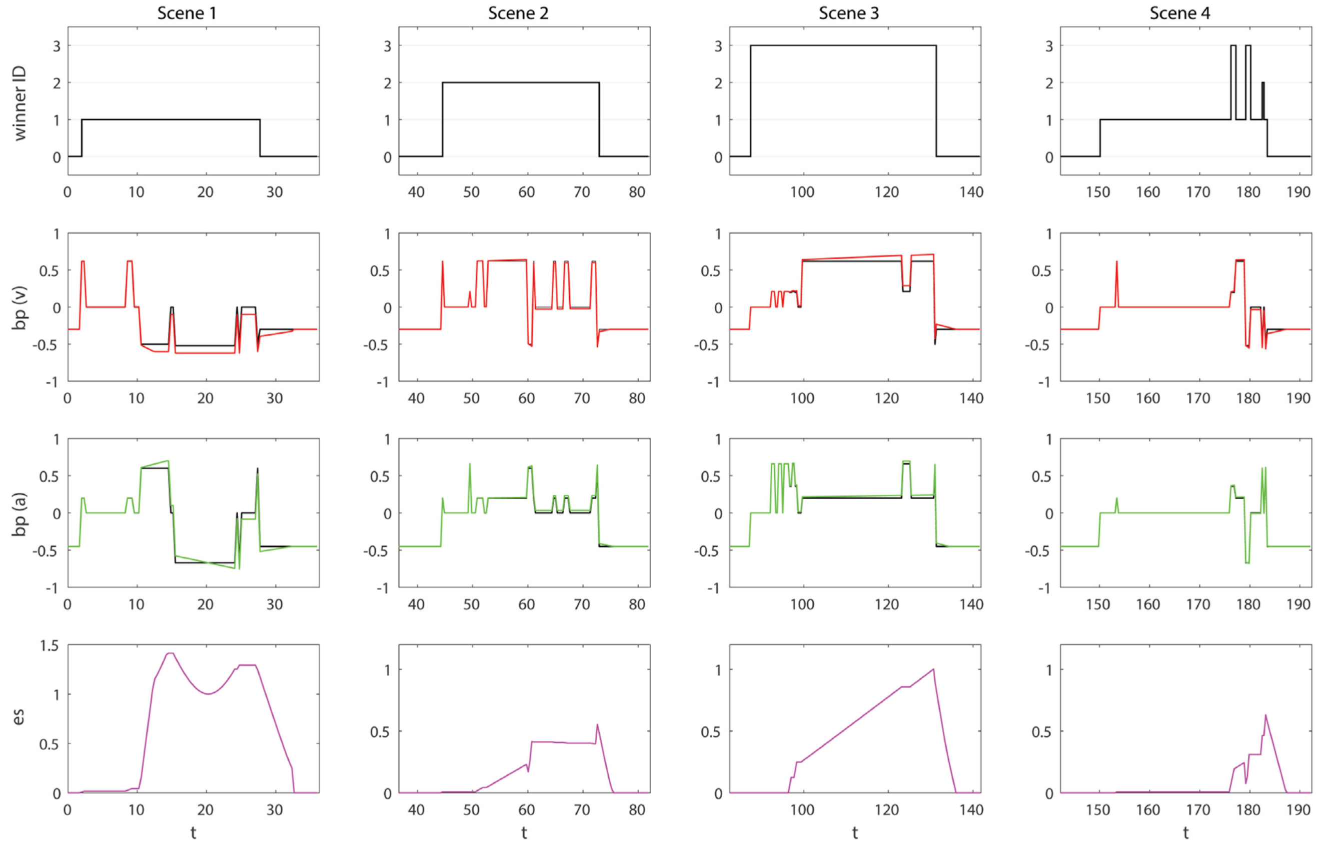


FIGURE 9 | Results of the HRI experiment with FACE integrating SEAI in condition 2. Columns are the four scenes. Rows are in order: *winner ID*, *bp (v)*, *bp (a)*, and *es*. Effectively executed v' and a' are, respectively, colored as red and green lines, while *bp* values not affected by *es* are represented as black lines to allow comparison. Time *t* is expressed in seconds.

a_{bp}) values, but this emotional immediate reaction is modulated by the new internal representation of emotions. These new values are (v',a'), discussed in section 5, where the *influence factor* has been set as $k = 0.1$. The higher priority of ALT-BEHRS guarantees that (v',a') are executed instead of (v_{bp}, a_{bp}). In the charts of `bp(v)` and `bp(a)`, we report both the values with (red line for valence, green for arousal) and without (the black line underneath) the `es` contribute. As expected, their difference is proportional to the intensity and the duration of the emotional state perceived. Moreover, the trend of `es` is slower and can last more than the duration of the causative stimulus, as in the transitions from detecting subjects to loneliness, which is no more immediate but smoothed (e.g., scene 3, `es` and `bp` after $t = 130$).

At this stage, the agent is aware of its own simulated feelings thanks to a continuous assertion of facts in its working memory reporting its own synthetic emotional state. Feelings also emerge in the body as shades of the emotional states expressed by the agent. In any case, all this information is temporary, there is a modulation of the behavior but still, no clear connection between the causative stimulus, the agent body state, and the subsequent feeling perceived. As a consequence, a recall of emotions driving specific behaviors is not feasible and the deliberative behavior of the agent is approximately the same: ID1 is still the winner of FACE attention.

6.1.3. Experiment 3

The addition of *FOF module* results in the definition of SOMARS and the possibility for SEAI to exploit the somatic marker mechanism. In **Figure 10**, we can see the results of the experiment in this third condition. The difference is impressive: during the first three scenes, in which the agent interacts individually with the three subjects, the attentive behavior of the robot is exactly the same, but the emotions evolve in a very different way; while, in scene 4, in front of all the subjects the attentive behavior is completely changed, emotional reactions are more stable, and the emotional state perceived is zero. This is due to the SM creation and recall mechanism discussed in section 5.6. Referring to the experiment, sensibility has been set to $s = 0.75$, so, the annoying behavior of ID1 makes the `es` intensity increase rapidly until it exceeds the s threshold ($t = 15.5$ s), this leads, in the next run cycle ($t = 15.83$ s), to the creation of a SM containing the *winner ID*, a marker *value* of −74.4 according to the equation reported in section 5.6, and the current *bp (v,a)* induced by the causative entity. The same thing is happening when FACE interacts with ID3 during scene 3, but here the quality of the marker is positive (details in **Figure 10**). As soon as these markers are created, the emotional state is no longer perturbed by the marked entity, because the agent has a precise belief and an associated emotional behavior to express toward

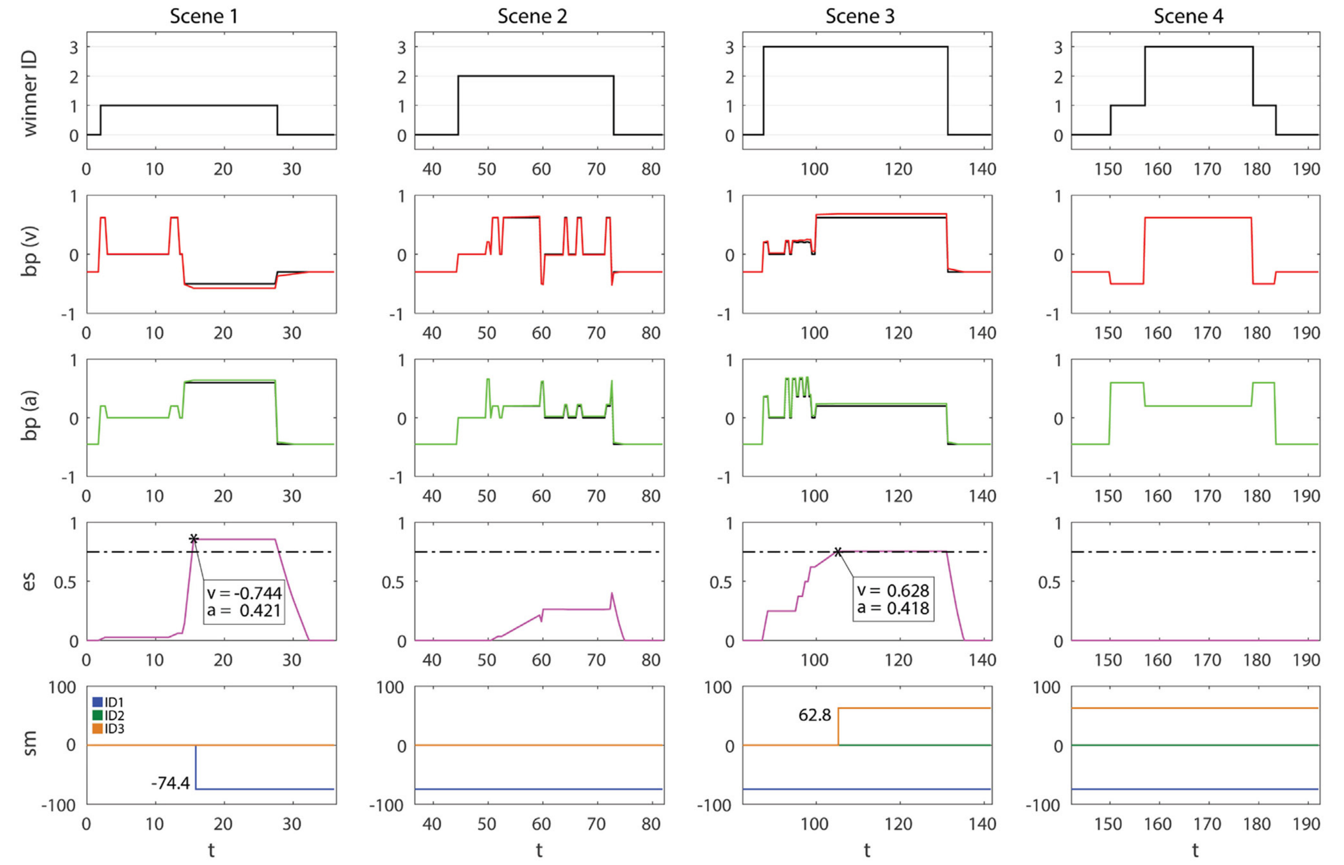


FIGURE 10 | Results of the HRI experiment with FACE integrating SEAI in condition 3. Columns are the four scenes. Rows are, in order: *winner ID*, *bp (v)*, *bp (a)*, *es*, and *sm*. Effectively executed *v′* and *a′* are, respectively, colored as red and green lines, while *bp* values not affected by *es* are represented as black lines to allow comparison. In *es*, we pointed out the *es (v,a)* values that caused the creation of a *sm*. In *sm* we reported the *values* of SMs in the moment in which they have been created by the system. IDs colors are indicated in the *sm* chart. Time *t* is expressed in seconds.

that specific subject, which is the somatic state felt and labeled through the somatic marker mechanism. This can be seen both in scenes 1 and 3 after the creation of the SM, and, which is more important, in the last scene. Indeed, in scene 4 when all the subjects are in front of the FACE robot, FACE is no longer attracted by the presence of the nearest subject. On the contrary, the presence of marked subjects completely bias its behavior: ID1 now is labeled, and when he enters and becomes detected, the robot immediately recalls the somatic state (−0.5, 0.6) felt in the past causative interaction; the same happens as soon as ID3 comes into the FOV of the agent. In our behavioral model, SPEC-BEHRS related to positive marked entities have higher priorities on rules driving the attention on negative marked entities. Therefore, until ID1, ID2, and ID3 are all detected, the attention of the robot is all for ID3. FACE is specifically attracted by him, thanks to his previous nice behavior, and stares at him with a pleasant facial expression (0.2, 0.68). In this last scene, ID2 becomes quite invisible to the robot, because his neutral previous interaction has never pushed the emotional state over the sensibility threshold (as shown in the `es` trend of scene 3). That experience did not influence enough the robot to create a dedicated SM.

This last experiment represents the test of the full SEAI system configured as Damasio's theory simulator endowed with the somatic marker mechanism. At this stage, the agent is able to autonomously create long-term memory information about entities of it social environment. These memories are emotional memories and are perceived by means of the body. They can affect the somatic state of the agent in case of further interactions, and bias the behavior in a very evident way. This mechanism, completely bio-inspired, let the agent automatically build its own beliefs about the outer world and about itself. What has been described, to all intents and purposes, models the construction of an autobiographical emotional memory and it respects the minimum requirements for the emergence of what Damasio described as an *Extended Consciousness*.

7. DISCUSSION AND CONCLUSIONS

In this paper, a novel cognitive architecture for social robots has been presented. We selected a well-known mind theory to be modeled and implemented in the form of a cognitive system controlling an emotional robot with sophisticated expressive capabilities. The developed system is called SEAI (Social Emotional Artificial

Intelligence). In particular, it has been inspired by the findings of Antonio Damasio and it is consistent with the computational formalization made by Bosse et al. (2008). It is based on a declarative rule-based expert system on top of procedural services deputed to the perception and motion control of the robot. Compared to other robotic cognitive systems, some of which discussed in the state-of-the-art section, SEAI has still some shortages: homeostasis control is missing, the agent's physiological parameters are a symbolic representation, capabilities such as perspective-taking or mind-reading have been not yet considered. Most of the effort has been spent in the C1 meaning of consciousness, rather than in the C2 definition (Dehaene et al., 2017). On the other hand, SEAI stands out from the other systems thanks to the hybrid concept with which has been designed. Indeed, the modular design of the architecture potentially enables the extension and portability of the system to any other social robot simply adapting, or adding, low-level services to the sensory apparatus and the motor system of the specific agent. This can be done keeping the "personality," memories, beliefs, experience, and behavioral traits of the agent, all of which depend on the cognitive part of the system, and therefore can be transferred or modified independently. Moreover, the innate extensibility of the rule-based expert system, which is the core of the cognitive block, puts no specific limitations to the inference reasoning capabilities with which the artificial agent can be endowed, which depends on the number and complexity of the rules. In the presented experiments, SEAI endowed a social humanoid with artificial emotions and feelings that have been influenced by the context, the agent managed to exploit them to build opinions on the social world in which is immersed, and, based on them, it manifested more sophisticated social skills. For instance, in the last experiment, an evident bias from the robot's standard behavior emerged. Such experiment obviously does not pretend to be the demonstration that we created a conscious being, but it is a clear demonstration of how SEAI and the chosen "understanding by building" approach lead to an important confirmation: with SEAI, robots can benefit from their own artificial emotions for taking decisions and treasure their past interactions. Future works will include (1) the expansion of SEAI in order to include the missing features identified in the other robotic cognitive systems; (2) the simulation of many other complex human social behaviors by writing new rules and expanding the current rule- sets; (3) study of the people's reactions to the adaptation of the robot behavior to its social environment by means of HRI experiments, eventually on long-term interactions. For the purpose of points (2) and (3), the involvement of professional figures from behavioral psychology and neuroscience would be greatly fruitful, and a questionnaire investigating the interlocutors feedback about the perceived consciousness of the robot will be required. The key issue is if the social interaction with humans would effectively benefit from the created deviations in the behavior of the social robot. Our hypothesis to test is that the realism derived by the integration of SEAI will improve the acceptability and the believability of this new kind of robots. In conclusion, we believe that SEAI is a potential valuable tool for modeling human consciousness and, ultimately, a promising beginning to tackle the possibility to attribute to the robots a synthetic form of consciousness. In this latter case, ethical issues will become extremely relevant and critical.

ETHICS STATEMENT

The experiment of this study has been approved by the Ethics Committee of the University of Pisa (prot. 68459, ref. Ethical Approval by CEAVNO, Comitato Etico di Area Vasta Nord). All research participants provided written and informed consent.

AUTHOR CONTRIBUTIONS

LC is the main author of this paper, he formalized the Damasio's theory in order to port it on the SEAI system, wrote the rules of the expert system, and designed the experiments to test the SEAI efficacy; DM designed the software architecture at the core of the SEAI framework; DR is full professor of bio-engineering and robotics who drove the focus of the entire research proposing Damasio as consciousness model to be analyzed. DR also did a strong contribution to the organization, writing, and proofing of the presented paper.

ACKNOWLEDGMENTS

We thank our colleagues Roberto Garofalo, for his fundamental assistance during the design and conduction of the presented experiments, and Chiara Magliaro, who helps the authors in preparing a clean and elegant presentation of the results. We thank Franco, Alejandro, and Gianluca for participating in the HRI experiment.

FUNDING

This work was partially funded by grants from the European Research Council under the European Union's 7th Framework Programme within the Project EASEL (Expressive Agents for Symbiotic Education and Learning), grant agreement FP7-ICT-611971.

REFERENCES

Alemi, M., Meghdari, A., Ghanbarzadeh, A., Moghadam, L. J., and Ghanbarzadeh, A. (2014). "Effect of utilizing a humanoid robot as a therapy-assistant in reducing anger, anxiety, and depression," in *2014 2nd RSI/ISM International Conference on Robotics and Mechatronics, ICRoM 2014*, Tehran, 748–753.

Bartneck, C., and Forlizzi, J. (2004). "A design-centred framework for social human-robot interaction," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)* (Kurashiki: IEEE), 591–594.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cereb. Cortex* 10, 295–307. doi:10.1093/cercor/10.3.295

Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science* 275, 1293–1295. doi:10.1126/science.275.5304.1293

Berlin, M., Gray, J., Thomaz, A. L., and Breazeal, C. (2006). "Perspective taking: an organizing principle for learning in human-robot interaction," in *AAAI*, Vol. 2, 1444–1450.

Bosse, T., Jonker, C. M., and Treur, J. (2008). Formalisation of Damasio's theory of emotion, feeling and core consciousness. *Conscious. Cogn.* 17, 94–113. doi:10.1016/j.concog.2007.06.006

Bosse, T., Jonker, C. M., Van Der Meij, L., and Treur, J. (2005). "Leadsto: a language and environment for analysis of dynamics by simulation," in *German Conference on Multiagent System Technologies* (Koblenz: Springer), 165–178.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *Int. J. Hum. Comput. Stud.* 59, 119–155. doi:10.1016/S1071-5819(03)00018-1

Breazeal, C., and Scassellati, B. (1999). "How to build robots that make friends and influence people," in *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'99*, Vol. 2 (Kyonggju: IEEE), 858–863.

Breazeal, C. L. (2004). *Designing Sociable Robots*. Cambridge, London: MIT press.

Broekens, J., Heerink, M., and Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology* 8, 94–103. doi:10.4017/gt.2009.08.02.002.00

Brooks, R. A., Breazeal, C., Marjanović, M., Scassellati, B., and Williamson, M. M. (1999). "The cog project: building a humanoid robot," in *Lecture Notes in Computer Science*, Springer, 52–87.

Causo, A., Vo, G. T., Chen, I.-M., and Yeo, S. H. (2016). "Design of robots used as education companion and tutor," in *Robotics and Mechatronics*, eds S. Zeghloul, M. A. Laribi, and J.-P. Gazeau (Poitiers: Springer), 75–84.

Chella, A., and Manzotti, R. (2013). *Artificial Consciousness*. Andrews UK Limited.

Chung, W., Kim, G., and Kim, M. (2007). Development of the multi-functional indoor service robot PSR systems. *Auton. Robots* 22, 1–17. doi:10.1007/s10514-006-9001-z

Cominelli, L., Carbonaro, N., Mazzei, D., Garofalo, R., Tognetti, A., and De Rossi, D. (2017). A multimodal perception framework for users emotional state assessment in social robotics. *Future Internet* 9, 42. doi:10.3390/fi9030042

Cominelli, L., Mazzei, D., Pieroni, M., Zaraki, A., Garofalo, R., and De Rossi, D. (2015). "Damasio's somatic marker for social robotics: preliminary implementation and test," in *Biomimetic and Biohybrid Systems*, eds S. P. Wilson, P. F. M. J. Verschure, A. Mura, and T. J. Prescott (Barcelona: Springer), 316–328.

Crick, F., and Clark, J. (1994). The astonishing hypothesis. *J. Conscious Stud* 1, 10–16.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Grosset/Putnam.

Damasio, A. (2000). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Spektrum Der Wissenschaft, 104.

Dautenhahn, K., and Billard, A. (1999). "Bringing up robots or – the psychology of socially intelligent robots," in *Proceedings of the Third Annual Conference on Autonomous Agents – AGENTS '99* (New York, NY, USA: ACM Press), 366–367.

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science* 358, 486–492. doi:10.1126/science.aan8871

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown & Company.

Dennett, D. C. (1996). *Kinds of Minds: Toward an Understanding of Consciousness*. Basic Books.

Fernando, S., Collins, E. C., Duff, A., Moore, R. K., Verschure, P. F., and Prescott, T. J. (2014). "Optimising robot personalities for symbiotic interaction," in *Conference on Biomimetic and Biohybrid Systems* (Milan: Springer), 392–395.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Rob. Auton. Syst.* 42, 143–166. doi:10.1016/S0921-8890(02)00372-X

Giarratano, J. C., and Riley, G. (1998). *Expert Systems*. PWS Publishing Co.

Herlea, D. E., Jonker, C. M., Treur, J., and Wijngaards, N. J. (1999). "Specification of bahavioural requirements within compositional multi-agent system design," in *European Workshop on Modelling Autonomous Agents in a Multi-Agent World* (Valencia: Springer), 8–27.

Jonker, C. M., Treur, J., and Wijngaards, W. C. (2003). A temporal modelling environment for internally grounded beliefs, desires and intentions. *Cogn. Syst. Res.* 4, 191–210. doi:10.1016/S1389-0417(03)00004-4

Kidd, C. D., Taggart, W., and Turkle, S. (2006). "A sociable robot to encourage social interaction among the elderly," in *Proceedings 2006 IEEE International Conference on Robotics and Automation. ICRA 2006* (Orlando, FL: IEEE), 3972–3976.

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) the unconscious initiation of a freely voluntary act. *Brain* 106, 623–642. doi:10.1093/brain/106.3.623

Manzotti, R. (2006). An alternative view of conscious perception. *J. Conscious Stud* 13, 45–79.

Mazzei, D., Cominelli, L., Lazzeri, N., Zaraki, A., and De Rossi, D. (2014). "I-clips brain: a hybrid cognitive system for social robots," in *Biomimetic and Biohybrid Systems*, eds A. Duff, N. F. Lepora, A. Mura, T. J. Prescott, and P. F. M. J. Verschure (Milan: Springer), 213–224.

Mazzei, D., Lazzeri, N., Hanson, D., and De Rossi, D. (2012). "Hefes: an hybrid engine for facial expressions synthesis to control human-like androids and avatars," in *4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)* (Rome: IEEE), 195–200.

Metta, G., Fitzpatrick, P., and Natale, L. (2006). Yarp: yet another robot platform. *Int. J. Adv. Robot. Syst.* 3, 43–48. doi:10.5772/5761

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). The iCub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw.* 23, 1125–1134. doi:10.1016/j.neunet.2010.08.010

Pineau, J., Montemerlo, M., Pollack, M., Roy, N., and Thrun, S. (2003). Towards robotic assistants in nursing homes: challenges and results. *Rob. Auton. Syst.* 42, 271–281. doi:10.1016/S0921-8890(02)00381-0

Pioggia, G., Igliozzi, R., Ferro, M., Ahluwalia, A., Muratori, F., and De Rossi, D. (2005). An android for enhancing social skills and emotion recognition in people with autism. *IEEE Trans. Neural Syst. Rehabil. Eng.* 13, 507–515. doi:10.1109/TNSRE.2005.856076

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* 17, 715–734. doi:10.1017/S0954579405050340

Qureshi, F., Terzopoulos, D., and Gillett, R. (2004). "The cognitive controller: a hybrid, deliberative/reactive control architecture for autonomous robots," in *Innovations in Applied Artificial Intelligence*, eds B. Orchard, C. Yang, and M. Ali (Berlin: Springer), 1102–1111. doi:10.1007/978-3-540-24677-0_113

Russell, J. A. (1980). The circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi:10.1037/h0077714

Saerbeck, M., Schut, T., Bartneck, C., and Janse, M. D. (2010). "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, GA: ACM), 1613–1622.

Scassellati, B., Admoni, H., and Matarić, M. (2012). Robots for use in autism research. *Annu. Rev. Biomed. Eng.* 14, 275–294. doi:10.1146/annurev-bioeng-071811-150036

Seth, A. K. (2008). Causal networks in simulated neural systems. *Cogn. Neurodyn.* 2, 49–64. doi:10.1007/s11571-007-9031-z

Sharkey, A., and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics Inf. Technol.* 14, 27–40. doi:10.1007/s10676-010-9234-6

Vernon, D., Metta, G., and Sandini, G. (2007). "The iCub cognitive architecture: interactive development in a humanoid robot," in *IEEE 6th International Conference on Development and Learning, 2007. ICDL 2007* (London: IEEE), 122–127.

Verschure, P. F. (2012). Distributed adaptive control: a theory of the mind, brain, body nexus. *Biol. Inspired Cognit. Archit.* 1, 55–72. doi:10.1016/j.bica.2012.04.005

Vouloutsi, V., Blancas, M., Zucca, R., Omedas, P., Reidsma, D., Davison, D., et al. (2016). "Towards a synthetic tutor assistant: the easel project and its architecture," in *Conference on Biomimetic and Biohybrid Systems* (Edinburgh: Springer), 353–364.

Wada, K., Shibata, T., Saito, T., Sakamoto, K., and Tanie, K. (2005). "Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation. ICRA 2005* (Barcelona: IEEE), 2785–2790.

Webb, B. (2001). Can robots make good models of biological behaviour? *Behav. Brain Sci.* 24, 1033–1050. doi:10.1017/S0140525X01550128

Werry, I., Dautenhahn, K., Ogden, B., and Harwin, W. (2001). "Can social interaction skills be taught by a social agent? The role of a robotic mediator in autism therapy," in *Cognitive Technology: Instruments of Mind*, eds M. Beynon, C. L. Nehaniv, and K. Dautenhahn, Vol. 2117 (Berlin: Springer), 57–74. doi:10.1007/3-540-44617-6_6

Zaraki, A., Mazzei, D., Giuliani, M., and De Rossi, D. (2014). Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Trans. Hum. Mach. Syst.* 44, 157–168. doi:10.1109/THMS.2014.2303083

Zaraki, A., Pieroni, M., De Rossi, D., Mazzei, D., Garofalo, R., Cominelli, L., et al. (2017). Design and evaluation of a unique social perception system for human-robot interaction. *IEEE Trans. Cognit. Dev. Syst.* 9, 341–355. doi:10.1109/TCDS.2016.2598423

Conflict of Interest Statement: 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.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 08 March 2018
doi: 10.3389/frobt.2018.00021

The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition

Adam Linson[1,2,3]*, *Andy Clark*[4,5], *Subramanian Ramamoorthy*[6,7] *and Karl Friston*[8]

[1] Department of Computing Science and Mathematics, University of Stirling, Stirling, United Kingdom, [2] Department of Philosophy, University of Stirling, Stirling, United Kingdom, [3] Institute for Advanced Studies in the Humanities, University of Edinburgh, Edinburgh, United Kingdom, [4] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom, [5] Department of Philosophy, Macquarie University, Sydney, NSW, Australia, [6] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [7] Edinburgh Centre for Robotics, Edinburgh, United Kingdom, [8] The Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi
di Palermo, Italy

Reviewed by:
Tony Belpaeme,
Plymouth University,
United Kingdom
Owen E. Holland,
University of Essex,
United Kingdom

***Correspondence:**
Adam Linson
adam.linson@stir.ac.uk

Specialty section:
This article was submitted to
Humanoid Robotics,
a section of the journal
Frontiers in Robotics and AI

Received: 16 October 2017
Accepted: 16 February 2018
Published: 08 March 2018

Citation:
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

The emerging neurocomputational vision of humans as embodied, ecologically embedded, social agents—who shape and are shaped by their environment—offers a golden opportunity to revisit and revise ideas about the physical and information-theoretic underpinnings of life, mind, and consciousness itself. In particular, the active inference framework (AIF) makes it possible to bridge connections from computational neuroscience and robotics/AI to ecological psychology and phenomenology, revealing common underpinnings and overcoming key limitations. AIF opposes the mechanistic to the reductive, while staying fully grounded in a naturalistic and information-theoretic foundation, using the principle of free energy minimization. The latter provides a theoretical basis for a unified treatment of particles, organisms, and interactive machines, spanning from the inorganic to organic, non-life to life, and natural to artificial agents. We provide a brief introduction to AIF, then explore its implications for evolutionary theory, ecological psychology, embodied phenomenology, and robotics/AI research. We conclude the paper by considering implications for machine consciousness.

Keywords: free energy, uncertainty, self-organization, embodiment, evolution, affordances, skilled expertise, frame problem

1. OVERVIEW AND GENTLE INTRODUCTION TO THE ACTIVE INFERENCE FRAMEWORK (AIF)

In this article, we will consider the active inference framework (AIF)—or, more strictly speaking, the principle of free energy minimization (FEM)—as a principle, rather than as a hypothesis. This means that we do not consider evidence for or against AIF *per se*. As a principle, AIF cannot be falsified—it is just a formal description of dynamics (much like Hamilton's principle of least action; see below) that we apply to sentient agents. The process theories that attend AIF do, clearly, require evidence, which we refer to in our discussion.

Following a general overview, this section offers a gentle introduction to AIF, illustrating aspects of its instantiation as predictive processing (PP). Subsequent sections unpack the framework in greater detail, drawing out its implications for evolutionary theory, ecological psychology, embodied phenomenology, and robotics/AI research. In the final section, we directly consider aspects of machine consciousness.

Given the ill-defined concept of consciousness, we endeavor to bring onto the same page researchers from physics, biology, neuroscience, philosophy, cognitive science, and robotics/AI, by reviewing concepts that are sometimes presumed to have unique and self-evident meanings. This approach aims to dispel misinterpretations and sharpen the cross-disciplinary focus on the substance of the claims. Throughout the following exposition and argument, there are several deep connections to the possibility of machine consciousness, although this topic only emerges as central in the concluding section. The preliminary sections are a necessary prelude to appreciating the implications of AIF for biology and robotics/AI, given that the notion of consciousness in robotics is sourced from the human equivalent. It is, therefore, important to establish a perspective from which human biology is accounted for by a mechanistically grounded, information-theoretic treatment. This perspective can lend itself to robotic implementation; however, without this grounding, any arbitrary properties associated with consciousness could be thusly implemented, putting the proverbial cart before the horse in modeling the target phenomenon.

Embodied and embedded human cognition has been analyzed extensively, not only in cognitive science but also in ecological psychology and phenomenological philosophy. Furthermore, all three fields have continually engaged with robotics/AI, contributing insights and critical perspectives, in some cases even effecting technological shifts (see, e.g., Brooks, 1999; Dreyfus, 2007; see also Chemero and Turvey, 2007; Sahin et al., 2007). More recently, there has been a proliferation of fruitful exchanges between robotics/AI and neuroscience (Hassabis et al., 2017), especially with respect to PP. The generalization of PP in AIF makes it possible to bridge connections to ecological psychology and phenomenology, revealing common underpinnings and overcoming key limitations inherent to the latter two.

To indicate where this account is headed, our conclusion supports the idea that there is a fundamental relationship between (self-)consciousness and processual recursion, which has been suggested in other work (e.g., Maturana, 1995; Seth et al., 2006). To reach this conclusion, our discussion of consciousness is deferred throughout the paper, which tries to account for the emergence of processes and recursive architectures that underwrite a conscious embodied agent. In this light, we set up AIF in Section 1 in such a way as to be expanded upon in later sections. Sections 2 and 3 take a long view of the emergence of human biology that paves the way for the remaining sections. Sections 4 and 5 address relevant paradigm contrasts in computational treatments of perception and action, and their implications for both biological and robotics/AI research. Sections 6 and 7 explore theoretical implications and practical applications, concluding in Section 8 with a consideration of humanoid robot consciousness (the theme of this special issue).

1.1. Setting Up the Framework

AIF considers a thermodynamically open, embodied, and environmentally embedded agent (see, e.g., Friston, 2009, 2010; Friston et al., 2010, 2015a,b, 2016, 2017a,b,c). In AIF, the adaptive behavior of such a "cybernetic" agent is understood to be regulated by ecologically relevant information, underpinned by a perception/action loop. Taking a broad bio-evolutionary view, AIF regards the entire embodied agent as a generative model of the organism-relevant thermodynamics of its ecological niche (see below), in that the agent is a member of a phylogenetic species that is co-stabilized with its niche. This notion encompasses the reciprocal organism/niche coevolutionary relationship (Laland et al., 2017).

During later evolutionary periods in which organisms with neural systems arise, brains come to augment the more fundamental embodied agent with a neuronal-connectivity-based extension to the generative model that handles more complex organism/niche dynamics. Thus, even when discussing PP—the human (neuronal) instantiation of active inference—the brain should be understood as "taking a back seat" to the body, serving the body by facilitating more complex coordination. Such coordination, including the dramatic niche reshaping seen in human culture, serves to co-stabilize organism and niche.

For a bacterium or a plant considered as an agent (Calvo and Friston, 2017), the embodied biological inheritance (the stable species as generative model) can be regarded as an implicit, surprise minimizing, familiarity with the niche. Many (if not all) of the earliest species inherit all the mechanisms they need for responding to and reshaping their niche, to facilitate their own survival and development. Such brainless organisms should be kept in mind whenever we "skip ahead" to the AIF description of human neural architecture—and its role in navigating the complexity of our cultural niche.[1]

1.2. Generative Model Basics

We next introduce the core notion of a neuronally implemented generative model. Consider, for example, a first-time visit to a university campus. Since a university is a contingent cultural entity, no part of our biological inheritance should be expected to provide us with any campus familiarity. However, if we have any earlier exposure to other universities, from visiting, reading, or hearing about them, this experience may contribute to our expectations of familiar features: we could speculatively populate any given campus with some lecture halls, administrative buildings, cafes, and so on. This mental act of populating, in other words generating, amounts to using a *generative* model of a campus (i.e., generating consequences from causes). On a first-time campus visit, such a generative model allows us to "predict" (extrapolate from the model) that there is a cafe, or, more precisely, that there is a high probability of there being a cafe, even if in actuality, there is not one there.

If we are visiting a specific campus for the first time, our generative model will be rather vague, but as we gain familiarity, we fill in more details. This process of gaining familiarity is a form of exploration, which may entail wandering, reading signs, and talking to passers-by. The exploratory process amounts to updating or nuancing our generative model for

[1] For a related approach in philosophy of science, see, e.g., Bechtel (2014) and Bechtel and Abrahamsen (2007).

this particular campus, including specific buildings and their layout. The exploration fills in the blanks, so to speak, such that we can then exploit the model for explicit or implicit purposes, whether finding the shortest path to the cafe or aimlessly meandering on a leisurely stroll. If, when exploring the campus, every sensory impression evinces the right sort of predictions, you have effectively *inverted* your generative model. In other words, to update your model of *this* campus, it has to predict the right things in the right place at the right time. This process amounts to learning to recognize the causes "out there" in relation to their context-dependent sensory consequences, or more simply, getting a grip on how sensations are caused by attempting to predict them—and then learning how to predict in *this* context.

Thus, the explore/exploit dynamic in relation to a generative model of a niche (including any subset thereof) can be understood as a process of gaining familiarity and "leveraging" that familiarity to achieve any preferred outcome (Schwartenbeck et al., 2013). The generative model itself is augmented and developed through a broadly construed learning process that transforms neuronal networks. This developmental learning process throughout the lifespan is facilitated by, and supplements, the preceding evolutionary development of the embodied apparatus. Crucially, this learning entails something that gets quite close to conscious processing, namely a form of abductive inference that differs from standard accounts of perceptual inference, as we will see in later sections.

Significantly, in AIF, the gaining and leveraging of familiarity with respect to the generative model is not limited to agent-external (distal) phenomena. While seeing an apple in a tree is ordinarily thought of as perception (i.e., perceiving the apple or its qualities), AIF radically expands the notion of perception. In AIF, vision and the remaining four classical senses are part of exteroceptive perception, or exteroception. Beyond exteroception, however, motor-system-governed biomechanical actions, such as plucking an apple from a tree, can be perceived not only by exteroception (by sight and touch), but also by what is referred to as proprioception. Even in seemingly isolated vision, there is continuous interaction between extero- and proprioception, as visual sensing interacts with eyeball, head, and even whole-body movement. This is a fundamental move beyond PP *per se*; it acknowledges that simply making sense of sensory data is only half the problem. You also have to actively coordinate your sensory surfaces and, essentially, become the author of your own sensations. We will see later that the imperatives for the active sampling of the environment, subsequent inference, and consequent learning, all comply with the same imperative, namely to enhance familiarity or resolve uncertainty and surprise.

A further perceptual modality accounts for the sensing of hunger and related internal sensations that are not necessarily discernible through extero- or proprioception. These internal sensations are grouped together as interoception. Here, too, we must recognize the continuous interactions between interoception and the other modalities, whether in bacteria or humans. For bacteria, the generative model embodies continuous relationships between extero-, proprio-, and interoception in the form of chemotaxis and flagellar movements. For humans, when we feel an afternoon lull as a need for a snack, extero-, proprio-, and interoception interact, guiding us to the cafe to satisfy our hunger. In this light, the expanded notion of perception in AIF stretches well beyond the traditional sense of seeing the apple, in that it brings all perception and action under the same umbrella of ecologically embedded adaptive behavior.

1.3. Further Preliminaries

The full scope of the embodied (and optionally neuronally augmented) generative model in AIF includes the building and leveraging of familiarity with the array of interactions between extero-, proprio-, and interoception. This familiarity may be gained during the lifespan, as in human development, or it may be predominantly biologically inherited, as with bacteria. Across all cases, however, the agent seeks to bring about its preferred and familiar future (e.g., satisfying hunger) by advancing the state under its generative model, through a sequence that begins with its present state, and follows a pathway guided by (inherited or learned) familiarity. Given the exteroceptive dimension, the agent's state can always be more comprehensively understood as the joint state of the agent/environment system.

Despite the relative simplicity of the basis of AIF—an embodied generative model with interactive modalities that facilitate agent/environment state transitions—the framework elegantly scales up from bacteria and plants to humans, even in atypical cases: a caring individual who sacrifices their own life for a preferred or expected future in which someone they rescue survives; a psychedelic drug taker who seeks a perpetually exploratory series of wild hallucinations over a more stable experience; a prisoner on a principled hunger strike who attempts to bring about a future, not of sated hunger, but of some greater social justice. In all instances, agents are interactively reducing their uncertainty in an open-ended self/world relationship ("what will happen" or "what would happen if I did that").

This process of bringing about a preferred future is referred to (in AIF) as *active inference*, a concept that will be further fleshed out in the remaining sections. At present, it should already be clear why active inference is not continuous with earlier notions of perceptual inference, given the role of the three modalities accommodated by the generative model—especially when we consider that proprio- and interoceptive predictions change the sensory evidence for our percepts (*via* motor and autonomic reflexes, as we will see later). Arguably, even the AIF treatment of perception itself is not continuous with earlier theoretical treatments of perception, since in AIF, perception is deeply situated in the embedded context of the active agent. Moreover (as we will also see later), AIF goes beyond established paradigms critical of traditional perceptual inference such as ecological psychology, which, despite its action-oriented perspective, still exhibits a latent exteroceptive-centrism.

A final and highly significant meta-theoretical feature set of AIF—one that should appeal to humanities scholars who are wary of naturalistic and information-theoretical accounts of humanness—is that the framework inherently enshrines the fundamental uncertainty and unknowability of the future, along with the agent's fallibility about the present and past. In addition, in contrast to superficially similar accounts, AIF markedly opposes the mechanistic to the reductive. These features will emerge more clearly throughout the paper. The next section addresses the role of the free energy principle, "the other side of the coin" of active inference.

2. DEMYSTIFYING FEM: FROM PHYSICS TO INFORMATION THEORY AND BACK AGAIN

In this section, we use a version of Maxwell's "demon" thought experiment to illustrate how concepts such as entropy and equilibrium link thermodynamics and information/control theory in cybernetics (e.g., Ashby), especially regarding how this link pertains to self-organization and the regulation of coupled systems. Readers already familiar with these concepts may wish to skip this section. In Section 2.1, we provide an introductory account of statistical thermodynamics and associated concepts, such as FEM, entropy, and uncertainty. We then connect these concepts to information theory and cybernetic control theory in Section 2.2. Finally, in Section 2.3, we return to thermodynamics, with an emphasis on substrate limitations for physically realized computational process models.

2.1. Thermodynamic FEM, Entropy, and Uncertainty

It might seem far-fetched to think that the entire universe has a direct relationship with a personal computing device. And yet, from the standpoint of thermodynamics, your laptop heats up because of the work it is doing shunting around subatomic particles, which in turn directly increases the total entropy of the universe. Of course, cosmologists have little interest in the vanishingly insignificant impact of a laptop on the universe. Scale matters a great deal in thermodynamics, because any thermodynamic system is an artificially bounded subsystem of the universe, which by stipulation, resides at the largest end of the scale. In this sense, the timescale of the universe offers the longest possible temporal trajectory, into which all other system trajectories eventually collapse.

It is a theorem in physics that the total entropy of the universe continuously increases (a corollary of the second law of thermodynamics). Thus, for any subsystem, whether a galaxy, organism, or even a laptop, if it can in any way reduce entropy within its system boundaries, this will only be for a *relatively* short time[2] until it must yield to the entropy-increasing pressure of the universe. This relationship can be viewed as a process of maintaining a local state equilibrium at the temporary expense of a global state disequilibrium; the global state will eventually reclaim its equilibrium in the long run by overwhelming the local state.

Thermodynamic entropy can be understood as a measure of our ability to predict the position of particles within a system over a duration. This is why entropy typically increases with heat,[3] since generally speaking, faster particle movement gives off more heat than slower movement, and faster movement leads to more-difficult-to-predict positions. Conversely, cooling slows down particles, making their positions more predictable, thereby decreasing entropy. Another way to describe the predictability of particle positions is in terms of our relative certainty about their predicted positions (in relation to the limited set of all possible positions). In this sense, higher thermodynamic entropy, greater unpredictability, and greater uncertainty are all linked to the same underlying quantity.

To bring together the notions of equilibrium states and entropy, consider a modern refrigeration unit. Its interior is kept cool by the operation of an electrical motor that gives off heat outside the unit. The entropy of the room (and indeed the universe) that houses the unit, i.e., the global equilibrium state, increases by the operation of the motor, while the cool interior, i.e., the local equilibrium state, momentarily maintains a lower entropy than the exterior. Eventually, of course, over the long run, the motor will stop, finally rewarming the unit. For keeping our drinks cool, however, it suffices to focus on the local subsystem and its corresponding timescale.

Finally, we reach the notion of FEM. In thermodynamics, particle movements count as work, and work has two main energetic effects: it uses some energy to do the work, and it releases some energy as light and/or heat. The energy available or "free" for the work is, thus, un-mysteriously referred to as free energy, in contrast to the available energy already (lawfully) dedicated to being released during the work. Returning to the above example, in a room with a refrigerator, when the fan has warmed the room air, the warm air particles have sufficient free energy to expand across the entire room. As long as the refrigerator door is closed, those particles cannot penetrate the fridge, so they only expand to occupy the room minus the fridge (a disequilibrium between the global/room and local/fridge states). However, when the fridge door is opened, the warm air particles expend their free energy by expanding into the open fridge. In this sense, they (lawfully) minimize free energy, i.e., they use the available free energy to expand across the full space, including the fridge interior. That is, through thermodynamic FEM, the global equilibrium/high entropy state of the warm room overwhelms the local equilibrium/low entropy state of the cool fridge interior.

2.2. FEM, Entropy, and Uncertainty in Information Theory and Cybernetics

Imagine that when we open our fridge door, a tiny demon[4] appears, to swat away the incoming warm air particles. If it swats

[2] This, of course, could be millions of years.

[3] We specify "typically" here as a nod to the Fluctuation Theorem (that generalizes the second law to non-equilibrium systems). In brief, the Fluctuation Theorem says that the probability of entropy decreasing vanishes as the observation time or size of the system increases (Evans and Searles, 2002). In other words, at a microscopic level, it is possible to have transient decreases in entropy, but the probability of this occurrence quickly becomes almost zero, over time.

[4] Maxwell's demon is a thought experiment proposed by James Clerk Maxwell to account for violations of the Second Law of Thermodynamics (Maxwell, 1871, pp. 308ff.). Subsequently, it was realized that even Maxwell's demon complies with the Second Law in virtue of Landauer's principle, namely, that "any logically irreversible

away a few particles at a time, it can delay the inevitable process of the fridge warming up. The more particles it can swat away, the more prolonged the delay. Better still, what if it could swat away *all* incoming particles? This would be as good as leaving the fridge door closed, as the local equilibrium of the cool interior would be maintained (at least over the short run); anything less, and the global equilibrium state (the warm room) would overwhelm the cool fridge and spoil the milk.

This demon scenario illustrates what cybernetics pioneer W. Ross Ashby (1958) termed "the law of requisite variety." Requisite variety refers to the sufficient available responses by the local subsystem to resist the global system, such as the demon's sufficient responses to all incoming warm air particles to maintain the cool fridge. Without requisite variety, the global equilibrium is permitted to prevail in the short run.

Now imagine the demon is working as a remote operator, controlling the positions of the cool air particles in the fridge, and maneuvering them along the plane of the door-opening to block any incoming warm air particles. This leads the particles to bounce off each other while remaining on their respective original sides of the opening, in which case the local subsystem remains thermodynamically identical before and after the onslaught of repelled particles. Significantly, the *average* thermodynamic state of the entire local subsystem is not concerned with a subset of specific particle positions. And yet, in our example, it is precisely this subset of particle positions that serve to maintain the local equilibrium. In this respect, while differing particle positions can result in thermodynamically equivalent systems, the systems would be informationally distinct, in that they reflect different organizations of the same set of particles. This brings us to Shannon (1948) information theory.

For Shannon, the distinct informational notion of entropy is borrowed from thermodynamics, as suggested by John von Neumann, who noticed the affinity between the concepts (Levine and Tribus, 1978). Shannon recognized that a set of binary switches has many possible on/off positions that can, by stipulation, be assigned any meaning. When transmitting a set of positions as a signal over a channel, noise made up of the same elements of the signal increases along the length of the channel. As this noise increases, it clouds the source signal, which in turn must be distinguished from an increasingly greater set of possible on/off switch configurations. In this sense, the location of the signal in the noise becomes increasingly uncertain.

As with particle positions in thermodynamics, the greater the ability to "predict" where the signal is within the noise, the greater the certainty. Thus, informational FEM is a reduction of uncertainty, i.e., an increased probability of picking out the relevant signal from the noise. By analogy to physics, this quantified uncertainty is termed Shannon entropy. Higher Shannon entropy reflects a greater uncertainty in picking out the relevant information, so informational FEM amounts to improving the identification of the relevant information. Technically, Shannon entropy is the expected self-information (a.k.a. *surprisal*) that (variational[5]) free energy aspires to approximate. This means that if one minimizes variational free energy at every point in time, the time average or expected surprisal is likewise minimized, thereby minimizing Shannon entropy *via* FEM.

Since the signal for Shannon is merely a particular organization of a subset of the same elements comprising the noise, the organization itself constitutes the relevant information. Of course, different organizations of the same source may be relevant under different circumstances. In Section 6.3, we will consider this sense of variable relevance in relation to the frame problem. Here, we focus on a narrow sense of relevance that builds on Ashby's law of requisite variety.

Conant and Ashby (1970) introduced the Good Regulator Theorem. This holds that, when two systems are coupled, given requisite variety (as with our demon controller), one system can remain in its local equilibrium state (cool fridge interior), despite the pressure of the system in a global equilibrium state (warm room). Without requisite variety, the system with greater variety will overwhelm the other, subsuming it into the global equilibrium. Requisite variety can be thought of a system having sufficient control information—and response parameters—to maintain its local equilibrium (the demon re-organizing the particles). In this sense, the system is a "good regulator" of the global system and on this basis, behaves as a model of the global system. We will see later that this translates into an agent with the right sort of generative model that can generate the consequences of a variety of actions.

Crucially, using this theorem, Shannon entropy can be transformed into a sender-free construct. Specifically, for the model in local equilibrium resisting the global state, it must not only have sufficient parameters, but it must pick out the "correct" organization of elements from the global system (such that "correct" refers to the information that allows the local system to resist being overwhelmed). To illustrate the sender-free notion of Shannon entropy with the fridge example, note that there is high uncertainty concerning which subset of warm air particles and their positions will threaten the open fridge door boundary. If the demon does not continuously select and re-organize the interior particles into the "correct" (blocking) positions, the milk spoils. Informational FEM amounts to the reduction of uncertainty (sender-free Shannon entropy) concerning the warm air particles, without there being a sender transmission *per se*. This will be important later (to Gibsonians, among others) for understanding that, on the AIF conception, the environment does not *transmit* information to the ostensible sensory-receiver.

2.3. Design Requirements for a Brain

Finally, we return to thermodynamics, in a slightly different role. Imagine replacing our demon with an ordinary laptop running special software to perform the same role described above (identifying and blocking incoming warm air particles), with one

manipulation of information, such as the erasure of a bit or the merging of two computation paths, must be accompanied by a corresponding entropy increase in non-information-bearing degrees of freedom of the information-processing apparatus or its environment" (Bennett, 2003).

[5] We will use the term of variational free energy (in information theory and Bayesian statistics) to distinguish it from thermodynamic free energy in FEM.

additional constraint: the laptop must be placed inside the fridge. Lacking the demon's thermodynamic law-defying properties, the laptop emits heat whenever it computes and controls the particle organizations. Thus, it is potentially self-defeating, since it threatens to raise the interior temperature despite keeping the outside forces at bay. Engineers could in principle redesign and reprogram the laptop to achieve efficient blocking by performing relatively few computations. A poor design might run too hot or too unreliable to be useful, while an ideal design would not overheat and block just enough particles to keep the milk cool.

This is why it is not enough to say that a thermodynamic system at local equilibrium can be a good regulator of a greater system by informational FEM alone. The local system must do thermodynamic work to be a good regulator of the greater system.[6] Thus, the local system architecture must accomplish this work without a self-defeating heat increase (which would also increase thermodynamic entropy). This points to the fact that the means by which informational free energy is minimized must simultaneously serve to minimize thermodynamic free energy in order for the local system to maintain its equilibrium. We will see later that this theme is central to notions of efficiency, simplicity, and the elimination of redundancy that is inherent in FEM.

3. EVOLUTION THROUGH A CYBERNETIC LENS: SELF-ORGANIZING SYSTEMS, EMBODIMENT, AND ECOLOGICAL ADAPTATION

Building on the previous section, we show how FEM can be used to make sense of self-organization and embodiment. We first show how physical chemistry models build on statistical thermodynamics, and how biological models build on a chemical conception of metabolic processes. We then show why physical and informational requirements are relevant to understanding embodied biological agents in relation to the coevolutionary development of species and their ecological niches.

3.1. Self-Organization and System Boundaries

The multiscale self-similarity of thermodynamic FEM comes into clear focus in physical chemistry. In a chemical system, predicting the behavior of individual particles can be intractable, but we can use the same mathematical models for particle aggregations as for individual particles. A transparent example of this is the process of crystal formation, called nucleation (Auer and Frenkel, 2001). In a pool of solute, many particles are distributed throughout. Typically, the behavior of the liquid is such that, for the particles to minimize (thermodynamic) free energy, they simply follow the liquid flow patterns (i.e., the paths of least resistance, in other words, the least surprising trajectories). However, if the right subset of particles comes into proximity, their thermodynamic FEM will in fact lead them to aggregate together. This particle aggregation will continue to swirl around in the pool and, at various points, more particles will begin to follow a pathway that affords greater FEM by joining the aggregation than by swirling around apart from it. The aggregation becomes the nucleus of an emergent crystal formation, which reaches a critical tipping point that leads an increasing number of particles to join up with it in a crystalline structural arrangement—all this mandated by simply following the path of least resistance at each point in time.

In virtue of this pattern, the crystal is distinct from the pool: it is an emergent self-organizing system with sharp boundaries. Specifically, the crystal is a free-energy-minimized molecular arrangement which has a lower-entropy local equilibrium than the contrasting higher-entropy global equilibrium of the pool. Of course, the crystal is merely an inanimate rock. Consider, however, another equivalent self-organizing criticality system, a forest fire (Drossel and Schwabl, 1992; Malamud et al., 1998). There is a critical tipping point at which the chemical process of the fire gains the capacity to spread according to a pattern of available fuel, to continue the chemical catalytic process. The forest fire, like the crystal, has clear system boundaries that emerge. Unlike the crystal, however, the nature of the fire's metabolic process means its system boundaries will not be maintained without additional fuel, in which case the fire will "die out."

This metaphor of fire "dying" aptly reflects the fact that biological systems also exhibit self-organized criticality, with a parallel metabolism that demands fuel to maintain system boundaries. A bacterium must obtain fuel from beyond its system boundaries to burn within those boundaries, in order to maintain them. Hence, there is a direct continuity and self-similarity across self-organizing aggregations-as-embodied systems from physics to chemistry to biology (Sengupta et al., 2013; Friston et al., 2015a,b; cf. Chemero, 2008; Bruineberg and Rietveld, 2014).

3.2. Ecological Context

At the biological level of description, the theoretical vantage point of ecology becomes relevant to understanding how organisms keep a positive balance in their metabolic bank account, so to speak. The cybernetic evolutionary lens described above reveals the connection of the embodied organism to the AIF notion of a generative model. Specifically, the embodied agent has a "do or die" to-do list to maintain its system boundaries, or more comprehensively, to survive and thrive. This list includes the agent obtaining fuel from its niche (to sustain its metabolism), avoiding active existential threats (e.g., predators), and also remaining within its embodied-apparatus-relative niche boundaries by not being a fish out of water, a land mammal falling down a ravine, or indeed any organism exceeding atmospheric thresholds of high and low temperatures and surface pressures.

Broadly, this set of agentive processes can be understood as an active engagement in a homeostasis/allostasis dynamic (Pezzulo et al., 2015), which more broadly still, can be regarded as adaptive behavior. For adaptive behavior to succeed, that is, for the organism to survive and thrive, it must have inborn and/or acquired familiarity with itself and its niche. In other words, the agent must be able to act on control information concerning its self/niche relationship (Friston, 2014). This control information can

[6] Note that this is an instance of Landauer's principle described in Footnote 4, speaking to the fact that there is no free lunch when it comes to trading information for energy—in any process, the two are essentially the same.

be understood as embodied system-boundary-internal adaptive behavioral guidance information, with the sole requirement that it is good enough for facilitating the agent's ability to survive and thrive, akin to satisficing (Simon, 1957).

Notice, however, that despite foregrounding the importance of boundaries, the picture is one in which living organizations are themselves changeable in ways that minimize the free energy of an evolving process (see, e.g., Clark, 2017). Notice also that, despite the sometimes-grim connotations of cybernetics and control theory, the notion of "control" is here synonymous with regulation, in the sense that you control, i.e., regulate, your own appetite simply by eating. In this sense, for the organism to be a good regulator, it must have a satisficing degree of certainty about itself and its niche to pick out what is relevant to its "to-do" list, such as responding to perceived hunger or danger, e.g., by seeking food or shelter. In logically equivalent terms, the agent must reduce its uncertainty, i.e., minimize (variational) free energy for a thermodynamic payoff.

To achieve this FEM, on an evolutionary timescale, organisms may mutate and potentially become an embodied generative model of a new niche. On a lifespan timescale, they may explore their niche to learn its contours, find new sources of sustenance and shelter, and new threats to avoid, i.e., augment their inborn generative model. In the interplay of evolutionary and lifespan trajectories, organisms transform their niches, bringing about higher-certainty correspondences to some aspects of their embodied generative model (e.g., tunneling underground to cushion light sensitivity). Indeed, some perspectives in theoretical biology speak to evolution itself as a FEM process, for instance, generalizing Darwinian processes as physical implementations of Bayesian inference (Frank, 2012; Lammert et al., 2012; Campbell, 2016).

Early lineages of organisms including bacteria and plants respond to self and environmental regularities even without a neural system, whereas later lineages including humans have the further support of a neural system to respond to more statistically complex regularities. Such complexity is reflected by increasing neuronal connectivity throughout the evolution of stable species. The ability to identify regularities in control information that reflect (self and niche) thermodynamic regularities can thus be viewed as an ecological adaptation requirement. By attaining effectively low uncertainty concerning adaptively relevant niche information—that is, by continuously minimizing (variational) free energy—the embodied agent is able to maintain a stable local (thermodynamic) equilibrium. The agent thereby resists the potentially overwhelming pressures of the environmental global equilibrium (the second law of thermodynamics) for the limited duration of its lifespan.

3.3. Complexity and Spatiotemporal Integration

Given our account thus far, it should be clear why, from a "good regulator" perspective, the more informationally complex the niche, the more complex the embodied (and eventually brain-augmented) generative model must be to facilitate effective adaptive behavior. The basic reflexive behavior, from bacterial chemotaxis to some plant and even insect behaviors, indicates that the preponderance of adaptive "work" can be done at a deeply embodied level, with low-level connectivity requirements (see, e.g., Mann et al., 2017). This is why for Gibsonian ecological psychology and Brooksian robotics, the bulk of relevant regularities are regarded as being wholly external to the embodied (natural or artificial) agent.

However, the theoretical framing device positing that "the world is its own best model" (Brooks, 1999) ultimately does not scale up to account for more complex agent/niche interaction dynamics. From the AIF perspective, it might be said simply that the world is its own best *world*, while the embodied agent itself is the best model of those aspects of the world relevant to its surviving and thriving—a familiar econiche that it has largely constructed for itself (Laland et al., 2017). Arguably, in relation to evolutionary natural selection pressure arising from niche saturation, mutants will only survive to stabilize as a new species under one of two conditions: expanding into a new niche that is spatially beyond the saturated niche, or expanding into one that is spatially coextensive with it, but presents a different set of relevant regularities (see Ito and Ikegami, 2006). In the latter case, the corresponding increasing informational complexity of the niche plausibly relates to increasing organismic complexity (coevolution).[7] Once neural systems emerge, this coevolutionary pattern continues with increasing neuronal connectivity (Yaeger, 2009; see also Seth and Edelman, 2004; Yaeger and Sporns, 2006; Yaeger, 2013).

Continuing with this account, a significant meta-theoretical feature of AIF can be noted, namely, that the human individual is re-contextualized as emerging naturally from the social group. There has been increasing interest in socially grounded neuroscience (e.g., Dumas et al., 2010; Dumas, 2011) and social robotics (Leite et al., 2013). Yet, some accounts largely consistent with AIF (e.g., Butz, 2016) only consider the social as an afterthought to the individual. Under the above considerations, however, given the upper bound on individual brain capabilities from a thermodynamic perspective, for humans to stabilize as a species, social cooperation offers the greatest advantage for establishing an adequate niche to sustain a stable population (see Yoshida et al., 2008). Indeed, identifying evolutionary stable strategies in multi-agent games, within AIF, can lead to some counter-intuitive yet compelling conclusions, particularly in terms of the degree of sophistication agents require in relation to others (see Devaine et al., 2014).

At the same time, as human culture emerges, introducing even greater niche complexity, the very same cooperative distributed information dynamics can lead to inherent difficulties. It is intrinsic to the underlying mathematical model of AIF that an apparatus which evolved for reducing uncertainty is equally sufficient for *increasing* uncertainty under particular circumstances. This is evident in social misunderstandings, such as mistaking the attributed motivation of a facial expression (Clark,

[7] Note that we are again appealing to the good regulator theorem. In other words, there is a homology between the complexity of the world being regulated and the good regulator that must embody a model of that world.

2015b, Section 2.9). The potential for the system to backfire, so to speak, is a consequence of the fact that human niche complexity includes social and cultural relationships, artifacts, language, and so on, which corresponds to substantially more complex neuronal connectivity in humans as compared to our evolutionary predecessors (Street et al., 2017). Even within human groups, a narrower, more predominantly physical, interpersonal local niche engagement (e.g., a stag hunt) requires considerably less informational complexity than the vast distributed neural/environmental information dynamics across a broad integrated physical and sociocultural niche. In the latter, agents face a greater challenge in leveraging more radically limited partial information (Ramamoorthy et al., 2012).

As neural complexity increases on an evolutionary timescale, the AIF model of the neural architecture is described in terms of an increasing number of interconnected hierarchical layers. These layers facilitate more extended spatiotemporal integration, with a growing set of nested local scales of time and space, ranging from the immediacy of the reflex arc, to ecologically situated behavior, to the lifespan. For instance, a beaver building a dam must be able to handle more extended time and space than a bacterium. Primates (including humans) exhibit nested spatiotemporal integration when interactively engaged in a dynamic situation or observing a visual sequence, as do humans when following along with speech or writing by integrating syllables into words, words into sentences, and sentences into a narrative (Hasson et al., 2008; Kiebel et al., 2008; Chen et al., 2015; Friston et al., 2017c; Yeshurun et al., 2017). This complex nesting, which has been implemented in robotics (Modayil et al., 2014), corresponds to a neural architecture that instantiates active inference in humans as PP, with growing empirical evidence of neurobiological substrate correspondences (Friston and Buzsáki, 2016; see also Clark, 2013, 2015b).

4. UNVEILING THE WORLD, UPENDING THE INPUT/OUTPUT MODEL OF PERCEPTION (AND ACTION)

With a focus on brains, this section shows how AIF upends the input/output model of perception (and action) still prevalent in embodied cognition and ecological psychology research, and perhaps even more prominently so in robotics/AI. As the full implications of this upending unfold, two major theoretical problems—the inverse problem and the frame problem—are revealed to be artifacts of the input/output model, such that AIF does not merely solve, but in fact dissolves these problems. Moreover, the philosophical concern raised against PP (and by extension, AIF), namely, that it entails or implies a solipsistic agent, hermetically sealed off from the world by an evidentiary boundary (or "veil"), is shown to be unfounded.

4.1. The Poverty of Indirect and Direct Perception

Is the embodied generative model stuck behind an "evidentiary boundary" (or "veil"), with no direct access to an outer world that is merely inferred? This is the notion of indirect perception that Hohwy (2013, 2016) advocates (cf. Clark, 2016). What Hohwy misses is a relevant distinction between phenomenal sensation and control information (elaborated in this section). Following the AIF account outlined above, control information provides the possibility for the agent being a good regulator. However, this remains distinct from phenomenal sensation of the world. At the same time, phenomenal sensation can itself be harvested for control information, in addition to information beneath the awareness threshold (Kang et al., 2017).[8]

Consider, for example, a video conference call apparatus. In an efficient design, the data flowing from one call participant to another will serve two simultaneous roles: a qualitative (content-relevant) role, in that the data underpin the audiovisual streams by which the parties can converse; and, at the same time, the data will serve a quantitative (content-irrelevant) role as control information, in that the data transfer rate will modulate the audiovisual resolution to compensate for bandwidth variation. In a parallel sense, in AIF, there is direct thermodynamic engagement between the agent's sensory surfaces and the world. This is precisely why we wear special glasses to view an eclipse, or earplugs at a loud concert: the direct engagement can be so powerful as to be biologically destructive. At lower intensities, light and sound contribute to a variety of enjoyable phenomenal sensations, and yet, they serve a dual role as control information. Under situations of acute existential threat, the control information may be the only relevant signal, whereas under presumed existential comfort (e.g., at the cinema), the control information may be largely dampened while (by cultural convention) phenomenal sensations are experienced for their own sake. Most quotidian cases lie somewhere in between these two extremes, such as eating to satisfy hunger while simultaneously savoring the sensory delights.

Given the broadly Helmholtzian inference tradition that Hohwy draws on, it is notable that this is precisely the kind of inference that Gibson (1979/1986) criticizes in his elaboration of ecological psychology, finding fault in theories in which "the outer world is deduced":

> The traditional theories of perception take it for granted that what we see now, present experience, is the sensory basis of our perception of the environment and that what we have seen up to now, past experience, is added to it (pp. 251ff.).

This critique motivates Gibson's positive account of "direct perception," also referred to as "information pickup" (Gibson, 1979/1986, pp. 147ff.). And yet, upon closer analysis, his positive account results in many of the same theoretical shortcomings as

[8] See Yahiro et al. (2017) for preliminary empirical support of this premise; their experimental findings point to different physiological pathways, e.g., low environmental temperature leading to involuntary shivering vs. the phenomenal sensation of coldness leading to voluntary warmth-seeking behavior. On the complex interplay between phenomenal sensation and preconscious information, see Sergent et al. (2013).

the inferential model he criticizes, as we will see below (cf. Fodor and Pylyshyn, 2002).

Both Helmholtz and Gibson ultimately inherit the same problems from the classical input/output model of perception. What Gibson criticizes in traditional inferential theories is the notion of passive input, which he replaces with active input—but it is still input! The active component in Gibson hints at the significance of proprioception, but ultimately, he assigns it an exteroceptive-centric role (Gibson, 1979/1986, p. 141). To make this argument, we first present the classical input/output model shared by computational perceptual theory (conventional in biology and robotics/AI) and contrast it with AIF.

4.2. Classical Computation vs. Active Inference

The classical input/output model of perception (and action) is the predominant model used in psychological, neuroscientific, and robotic explanations; this model also typically underlies the notion of neural computation and information processing, and it is ripe for retirement (Clark, 2014). AIF implies a vastly different conception of the relationship between perception, action, and the world, that also points to a different sense of computation and indeed perception itself. To understand AIF's ontological commitments and implications for perceptual theory generally, and for robotics/AI, we must examine the assumptions and implications of the predominant model.

The basic elements and processes of the classical/computational model can be generalized as follows: un-encoded ("raw") data from the environment ("world") is selectively sampled by the agent and encoded as input ("reading" the raw data). This raw data input, once encoded into the system, is then processed (beginning with "early perception"). This processing chain produces a decoded output, terminating as a percept (and potentially entering into a secondary stage related to concepts). After this discrete stage, as this story goes, an executive controller may then retrieve the percept (or concept) from storage and engage it in further action-relevant computations or reflexively issue a reactive action command.

Significantly, two major problems arise as mere artifacts of this model—the inverse problem and the frame problem. Both have given rise to countless accounts of how to bypass or solve them. Most famously, Marr (1982) produces a highly influential and elaborate account of how to solve the inverse problem, to get from the input stage to meaningful experience of the world. His solution comprises an elaborate series of "early" perceptual processing stages for disambiguating apparent equivalencies, implemented in subsequent decades of computer vision research. Marr was in part responding critically to Gibson's account, although some readings offer a middle ground between the two theories (Ullman, 1980; see also Shagrir, 2010). Gibson (1979/1986) and later analysts of ecological psychology argue that the inverse problem is bypassed without appealing to the kinds of processes Marr introduces (e.g., Hatfield, 2003; Chemero, 2009; Orlandi, 2017), for instance, by bodily movements (exploring or swaying) that reveal constant proportions in three-dimensional situatedness, in contrast to two-dimensional sources of optical projections. Like Marr, however, these ecological accounts still treat (what is regarded as) exteroceptive input as primary, even when the necessity of proprioceptive coupling is acknowledged.

Those who accept the classical/computational input/output model of perception must also face the frame problem (McCarthy and Hayes, 1969; Minsky, 1974), which can be generalized as a problem of knowing when and what raw sampling is needed for updating beliefs about the world (e.g., in relation to an isolated local action that only modifies a small subset of the environment[9]). It also concerns how to handle an input encoding from one context following a change of context. Thus, the frame problem is also known as the "relevance" (or "significance") problem, based on the premise that there is no obvious means of ascertaining what is cognitively relevant or significant under changing circumstances. The frame problem has led to elaborate logic-based solutions (Shanahan, 1997) and critical accounts of robotic AI based on embodied phenomenological philosophy (Dreyfus, 1992, 2007; cf. Wheeler, 2008).

4.3. Upending the Input/Output Model of Perception (and Action)

Building on the previous sections, we briefly show how AIF re-arranges the picture to dispense with the classical/computational model of input and output. Recall that above, we noted that there is direct thermodynamic engagement between the agent's sensory surfaces and the world, which requires protection from high intensities (e.g., earplugs at a loud concert). For an intuitive example of lower intensity engagement, consider a game of tennis. It would take some mental gymnastics to make sense of the idea that an arm is input to a racket, and a racket input to a ball—on this view, what would count as output? Instead, using basic physics, we regard the action of hitting the ball as a transfer of energy, from the arm to the racket to the ball. This same sense of thermodynamic energy transfer occurs between an organism's environmental niche and its sensory surfaces.

In AIF, the embodied agent learns the regularities of the sensory surface perturbations, much like what Gibson (1979/1986) refers to as invariants. Moving beyond Gibson, in AIF, the invariants extend across interactive regularities in extero-, proprio-, and interoception, in the form of the generative hierarchical model. The more regular covariance that is learned, such as how invariant proprioceptive hand-grasping patterns covary with invariant racket-swinging, ball-hitting patterns, the more reliable the generative model is as control information across a variety of conditions to which the model is adapted (see Kruschke, 2008). In PP, this adaptive process proceeds by a feedback loop with prediction error, i.e., minimizing prediction error amounts to adapting the generative model to the present conditions (Clark, 2013, 2015a,b).

The continuous embedding in the niche, which the agent explores to learn the covariance regularities, allows the agent to develop and update the generative model (akin to Gibson's notions of "tuning" and "resonance"). This goes beyond the exteroceptive-centric notion that minor proprioceptive alterations

[9] For discussion, see Sprevak (2005).

bypass the inverse problem. In AIF, the generative model links all reliably invariant information in a deeply situated way, such that perception and action enable the embodied agent to propel itself through a temporal succession of generative model modulations, for instance, approaching a distal food source to eventually alleviate hunger.

Under such situated embedding, the frame problem never presents itself, because the relevant aspects of the niche are thermodynamic perturbations, while engagement with the niche is facilitated by continuous control information. In the preponderance of ecologically valid conditions, there is never a temporally suspended slice of un-embedded input to be processed, nor is there an isolated (i.e., non-deeply situated) encounter with an exteroceptive input stimulus that is lightly probed through proprioception. That is, in real-world embodied and embedded cognition, there are no disconnected moments of perception of the world, since the world wholly envelops the agent throughout its lifespan. (We return to the frame problem in Section 6.3.)

Ambiguities arising from thermodynamically relevant niche details can indeed fail to be disambiguated, as they do during contrived experiments and illusions. However, in AIF, ambiguity is not an "early perception" input processing challenge, but rather a matter of the precision-weighting of layers of the hierarchical architecture (Friston, 2008). Many situated perceptual ambiguities can be accommodated by the precision-weighting of higher or lower layers: higher layers provide broad continuities to previous situations, such that ambiguities closer to the sensory surface can be ignored or recognized as illusory (as when the magician's assistant seems to disappear into thin air), while ambiguities at higher levels can be suspended pending further lower-level evidence (as when it is unclear if a friend entered the theater or joined the crowd outside). In addition, perceptual disambiguation is facilitated by the nested multiscale dynamics described above (Brascamp et al., 2008).

5. GIBSON RECONFIGURED: BEYOND RE-DESCRIPTION

Notably, AIF carries forward Gibson's core critique of his behaviorist and cognitivist predecessors; however, AIF also addresses the fundamental inadequacies of his positive account, as we illustrate in this section. We begin with an initial re-description or translation of some Gibsonian concepts into AIF. At relevant points throughout, we also highlight connections to robotics.

5.1. Initial Mappings

Recall from above Gibson's objection to theories (e.g., Helmholtz's) in which the present perception of the world is inferred by an additive process that uses the past (memory) to supplement missing details. Here, a technical clarification will be useful to distinguish traditional perceptual inference from AIF/PP. Shortly, we will flesh out what the actual process of "active inference" entails, but for now, it can be stated that in PP, the prediction of the present is fundamentally non-inferential in the traditional sense (see below for the specialized sense of surprisal-reducing model inference). Instead, perceiving the present is facilitated by an extrapolation from the environmentally embedded generative model. The model develops through biological inheritance and lifespan experience, based entirely on invariant covariance of modalities from past interactions.

Perception in AIF is thus not an additive process, but a generative one, which matters here for an important class of cases, namely, those in the cultural (as opposed to natural) domain. The cultural domain has physically bound cases with no natural equivalent, such as the operation of a door with a doorknob. We see many naturalistic examples in Gibson's writings, concerning, e.g., tunnels (which may occur in nature), but he also wishes to extend his theory to the human cultural environment (Gibson, 1966). Moreover, he wants to allow for a concept of learning (at best, coarsely defined), while simultaneously objecting to a model of mental storage and retrieval (Gibson, 1979/1986). How then, should it be possible to learn how a doorknob works such that "direct perception" of one (*via* ambient optical arrays) is at once the perception of a means for opening the door, without any specified mechanism for establishing this correspondence? If the correspondence is merely a conditioned association, then how can he avoid the claim (as he intends) that past experience is added to the present?

Despite Gibson's professed aversion to computation and traditional perceptual inference, the deeper problem here is that his theory recapitulates and is thus still bound by the classical/computational input/output model (cf. Bickhard and Richie, 1983). To better understand this issue, we must turn to his concept of affordances. For clarity, we will first establish how AIF re-describes aspects of Gibson's ecological framework in terms of the generative model.

In some AIF contexts (FitzGerald et al., 2014), it is more useful to treat the generative model as a model *space* populated with an ensemble of plausible generative models. For instance, consider a proprioceptive model of hand configurations: grasping, wrist rotation, peripersonal reach, and so on. To be clear, this sense of generative model is not an imagistic mental representation, but rather, a mathematical model of a set of invariant synaptic firing patterns that reliably correspond to bodily movements. These proprioceptive models (subsets of the complete generative model) are equivalent to Gibson's notion of organismic capacities. Within the model space, there are also exteroceptive models that reliably correspond to sensory perturbations caused by, e.g., trees and branches, doors and doorknobs, and so on, which relative to proprioception, re-describe Gibson's notion of environmental action opportunities (a branch affords climbing a tree, relative to the bodies of certain organisms). In his theory of affordances, Gibson also notes the relevance of the organism's wants and needs. These are incorporated into AIF as prior beliefs or preferences constituted by the generative model. Key among these are the priors over interoceptive predictions, by which we reliably come to recognize internal sensations such as hunger, fatigue, lack of fresh air, and so on (Seth et al., 2012).

Each of these models interact within a hierarchical model space, such that single modality invariants intersect and interact with each other, resulting in invariant covariance relationships: (interoceptive) hunger is reduced by eating fruit from a tree, which can be (exteroceptively) seen and (proprioceptively) reached by climbing branches. In a cultural context, the (interoceptive) need for fresh air can be met by (exteroceptively) transitioning from indoors to outdoors, as facilitated by a (proprioceptive) action sequence involving turning the doorknob and walking out of the room. The action sequence itself can be further broken down, in that even the doorknob interaction is a result of invariant covariance between exteroceptive control information and proprioceptive reaching, grasping, and turning; this principle has been successfully robotically simulated (Pio-Lopez et al., 2016). In brief, AIF offers a fundamentally embodied and embedded account of situated perception and action, rather than an exteroceptive-centric input/output model. The latter requires traditional perceptual inference based on early (perception) input processing of an impoverished stimulus; or, as Gibson has it, such inference is replaced by a woefully underspecified "direct perception" mechanism that fails to explain learned cultural affordances.

To summarize this initial re-description of Gibson's framework in AIF, and more importantly, the underlying shift in emphasis, we have seen that Gibson's affordances concern the perception of (a) environmentally specified information as action opportunities in relation to the organism's (b) embodied capacities and (c) needs and wants. In AIF, all three are integrated into the embodied (and neuronally augmented) hierarchical generative model, with correspondences to Gibson in terms of (a) exteroception, (b) proprioception, and (c) interoception. This allows us to make sense of a common ecologically valid scenario, such as the interoceptive need for fresh air, and the extero- and proprioceptive interactions that lead to turning the doorknob, opening the door, and walking outside. We are now in a position to flesh out what "active inference" itself refers to, which requires the introduction of a specialized concept: policies.

5.2. Affordances and Policies

The notion of policies highlights how the generative model can be temporally deployed over possible future states. Once this is understood, the full implications of embedded spatiotemporal nesting and its relationship to agent/environment dynamics can be brought into view. Policies are means of transitioning between states of the generative model, which can only be in one (actualized) state at a time.[10] The conventional sense of actions (e.g., reaching for the doorknob) "fall out" of policies, as we will see next.

A theoretician seeking to define a policy in propositional terms might define one (in the following example) as "go outside to get fresh air." The underpinnings of the policy are in effect a possible transition between two states of the generative model: the current state (at time t_0) and a preferred future state (at time t_1). At t_0, the agent is inside a room with a door to the outside. In the exteroceptive modality (in addition to phenomenal sensation), there is control information present concerning walls, doors, doorknob mechanisms, and so on. There is also proprioceptive (control) information available concerning, e.g., hand-grasping and leg-walking abilities. In the interoceptive modality, there is information concerning a sensed lack of fresh air and its presumed contribution to fatigue.

In this case, the preferred future outcome is having fatigue alleviated by getting fresh air. This would mean that if this outcome were attained, at t_1, the generative model would be altered, such that the exteroceptive information would pertain to an outdoor rather than indoor scene, and the interoceptive information would pertain to breathing fresh rather than stale air. To realize the preferred outcome, the agent *actively infers* the (t_0 to t_1 state transition) policy. Working backwards in a sense, to facilitate this transition, a series of actions "fall out," unfolding without requiring the planning of a sequence of action commands (Adams et al., 2013), in stark contrast to the robotics paradigm of sense-plan-act. Instead, the reliable covariance with proprioception and the other modalities of the generative model leads to reaching, grasping, and turning the doorknob, to open the door, to walk outside, to get fresh air, given that this set of covariances has been empirically established (i.e., learned).

The bottom line here is that if an agent entertains a generative model of the future, the agent must have beliefs (i.e., expectations) about future or counterfactual states under each allowable policy. Put simply, we have in mind here an agent whose generative model transcends the present and is continuously predicting the future (and past). Crucially, each prediction—at different times in the future—is subject to the same policy-dependent transition probabilities as apply to the here and now, thereby "connecting the dots" in a path to preferred and familiar outcomes. On this view, the present simply provides sensory evidence for one of several (counterfactual) paths into the future, where the path (or policy) with the greatest evidence gets to determine the next action. Notice again how we return to the path of least resistance or minimum (expected) free energy (i.e., maximizing model evidence over possible pathways).

Through a continuous series of perception/action loops, the embodied agent remains in open exchange with the world by actively probing its environment (Kruschke, 2008) and leveraging the control information of the generative model to alter the thermodynamic substrate (its physical position and condition). Even Gibson could not object to this sense of inference: there can be no "direct perception" of the future! Here, however, is where the uncertainty and unknowability of the future can be understood as a feature of AIF that is lacking in ecological psychology, namely, concerning *conditional* future outcomes. Even on the most charitable reading of Gibson, assuming we can explain (without magic) that one could "directly" perceive that "the doorknob affords opening the door" based on the ambient optical array, conventional affordance theory is left stranded in the face of an invisibly locked or broken doorknob. That is, when the doorknob fails to open the door, the exteroceptively ascertained ambient optical array remains identical before and after the attempt. Thus, within Gibson's framework, the doorknob forcibly remains an apparent affordance even with prior information that it does not open the door in this case. In such ecologically valid scenarios commonly faced by human cognition, it is a severe meta-theoretical weakness if they cannot be adequately addressed.

In contrast to ecological psychology, AIF elegantly handles conditional outcomes in terms of probabilities. This is why it uses

[10] Our description of active inference here will be based largely upon discrete time and state space generative models (e.g., Markov decision processes). These are simpler to handle in terms of their numerics (and possibly conceptually); however, the same principles apply to the continuous state space models usually considered in Bayesian filtering and predictive coding formulations of active inference.

a Bayesian model of neural processing, given that empirical priors derived from experience influence the generative model computations of probability,[11] a significantly different sense of computation than that used in input/output model descriptions (which hold that sampled input is computed/processed). Reconfigured by AIF, a typical affordance is merely a high likelihood, such that "affords" amounts to "offers a relatively sure bet." Thus, "the doorknob affords opening the door" is more accurately rendered as "the doorknob offers a relatively sure bet for opening the door," thereby accounting for the conditional outcomes in which the doorknob is locked or broken, unknowable by exteroception alone. In addition, when a source of information indicates a locked or broken state (such as a performed or observed attempt to open it, or by word of mouth), the doorknob ceases to be an apparent affordance, since it no longer offers the agent a relatively sure bet for opening the door, despite the fact that the ambient optical array is unaltered.

AIF is consistent with the view that "affordances are relations." More precisely, "affordances must belong to animal/environment systems, not just the environment," in that perceiving affordances is perceiving "the relation between the perceiver and the environment" (Chemero, 2003, pp. 185–6; see also Chemero, 2008). By adding the extended temporal dimension of AIF, the affordance relationality can be further understood as being between a presently given agent/environment relational state and probable future agent/environment relational states.

This move also allows AIF to account for conditions in a more distant future, such as dinner plans next week, which some theorists view as beyond the scope of ecological (and enactive) explanation. Here, such planning ability is seamlessly accounted for in the process of active inference. The plan sets into motion a series of intermediary interactions (actively inferred state transition policies) that propel the embodied agent toward the preferred future outcome. These interactions are based on experience and are, thus, deemed reliable (in a satisficing sense) with reasonably high probability, while (simultaneously) suggesting a low-probability capacity to fail. Put simply, all I need to do to determine my next action is to choose the most probable action under the prior belief: "I will not miss next week's dinner party." This prior belief generates a hierarchical cascade of empirical priors, each providing contextual guidance to accumulate the sensory evidence for the particular path I am pursuing. If everything goes well, this path would end successfully with arrival at the dinner party. Note that not only is there a deep generative model in relation to time in play here (Dehaene et al., 2015), there is also a hierarchical depth in terms of short and long-term policies, i.e., trajectories of states (see Friston et al., 2017c).

5.3. Free Energy, Revisited

What does all this have to do with the free energy principle? The policies the agent infers, as transitions from present to preferred future state, are those that minimize (variational) free energy expected on actualizing the preferred future state. This contextualizes the notion of reward motivations (that policies increase expected future reward) and even problem-solving itself, in that the reward or the solutions are part of the preferred future outcome as viewed from a present state (Friston et al., 2009, 2010; Friston, 2011; cf. Newell et al., 1959). Technically speaking, the expected free energy ensures that the prior probability of a policy maximizes reward (i.e., prior preferences) in the future, as in machine learning, under the constraint that it also minimizes uncertainty and ambiguity. Moreover, in the agent's relationship to the niche, expected free energy is minimized—uncertainty or disequilibrium is reduced (see Sections 2 and 3)—as the agent strives to select the relevant control information in the face of the densely rich informational environment (high Shannon entropy). This is an important point which takes affordances into the epistemic realm.

In other words, by trying to infer the FEM path of least resistance into the future (even for a challenging task), there is a necessary component of uncertainty that combines with prior preferences to determine the best policy. This means that the most probable policies or paths are those that resolve uncertainty when navigating the lived world (Berlyne, 1950; Schmidhuber, 2006; Baranes and Oudeyer, 2009; Still and Precup, 2012; Barto et al., 2013; Moulin and Souchay, 2015). To achieve this, agents engage in some interactions that serve an epistemic rather than pragmatic purpose, i.e., epistemic actions (Kirsh and Maglio, 1994). In AIF, we can place such epistemic actions in the general context of physical or mental epistemic foraging (Pezzulo, 2017), and further specify what facilitates such epistemic actions, namely, *epistemic affordances*. The latter concept brings with it the notions of salience—epistemic affordances that will reduce uncertainty about future states of the world—and novelty—epistemic affordances that will reduce uncertainty about the contingencies or parameters of my generative model. (The next section furthers this account of affordances.)

In summary, one's preferred future state is realized by exploiting high likelihoods in the sequence of state transitions of the generative model that underpins the agent/environment relationship (e.g., my relatively high certainty that my hand turns a doorknob, which opens a doorway, which I can walk through to get outside, to get fresh air, and to alleviate my fatigue). Exploiting high likelihoods refers to the probabilistic Bayesian decision-making computations that play out on a dynamic, neurobiological substrate (Pezzulo et al., 2015). In this context, it can be said that *local minima of uncertainty* (in the projected model state transitions) provide the critical points that can be leveraged to facilitate a preferred future (or avoid an undesired future). At the ecological "behavior" scale (policies), these local minima provide a comprehensive re-description of affordances that unites the exteroceptive with the proprio- and interoceptive dimensions (Pezzulo and Cisek, 2016). They also generalize to the sub-ecological "action" scale, as reflex arcs, grounded in the physics of nerve electricity (Friston et al., 2010; Sengupta et al., 2013), and the supra-ecological "activity" scale, as extended active and resting states, grounded in physiological homeostasis/allostasis dynamics (Ashourvan et al., 2017).

[11] See Albrecht et al. (2016) for an implemented reinforcement-learning-based decision-making model defined in terms of such probabilities (expectations).

6. SKATING UNCERTAINTY: GENERALIZED AFFORDANCE THEORY, SKILLED EXPERTISE, AND THE FRAME PROBLEM

This section considers how local minima of uncertainty in the projected temporal sequence of generative model states serve to unify developmental theory and the underspecified (by Gibson) notion of learned affordances. We then show concrete applications in skilled practical and cultural activities. Finally, drawing on robotics studies, we connect spatiotemporal nesting and agent/environment dynamics to adaptive policy reuse.

6.1. Generalized Affordance Theory

Here, we generalize affordances to every available reliable regularity in the agent/environment relationship, including basic objects. While this level of generality may seem meta-theoretically undesirable, it is worth bearing in mind that Gibson extended affordances to this high level of generality in explaining that air affords breathing, the ground affords standing on, cliffs are negative affordances for bipedal locomotion, and so on (Gibson, 1979/1986). On our account, affordances encompass the entirety of intuitive physics (see Clark, 2016).

As Franz and Triesch (2010) argue, a number of purported Gestalt percepts have only been considered in relatively late periods of individual (lifespan) human development, as even within the first several months after birth, there is a tremendous amount of densely rich environmental information encountered. The inborn apparatus (as suggested by AIF) for discerning regular covariance and leveraging that in situated activity can be computationally simulated with only a limited construct that yields a number of Gestalt-like phenomena. The limited construct—foreground and background differentiation—is a minimal mechanism that would be plausibly selected for on an evolutionary timescale.

In addition, there appears to be another plausibly selected for (inborn) minimal mechanism for differentiating inanimate from animate entities, with the latter possibly extending to finer-grained differentiations between conspecifics and other animals. There is evidence of this mechanism in brain scans of primates (Sliwa and Freiwald, 2017) and human infants (de Haan and Nelson, 1999, Southgate et al., 2008), and from human *in utero* behavioral experiments (Reid et al., 2017). This mechanism would plausibly underpin the fundamentality of social cooperation to human cognition (Barrett et al., 2010, Cortina and Liotti, 2010); a related point has been made about language, noting the fundamentality of dialog from which monolog is derived (Pickering and Garrod, 2004).

The above suggests that early developmental learning proceeds through interactive exploration (Stahl and Feigenson, 2015), which makes possible a high-level generative model of intuitive physics that augments inborn capacities with empirical priors. This is especially evident from the gradual development of coordinated bodily movement, ranging from basic crawling, walking, and stacking blocks, also explored in robotics (Pierce and Kuipers, 1997, Modayil and Kuipers, 2008, Ugur et al., 2011, 2012), all the way up to more elaborate activities such as interpersonally coordinated dancing and playing sports (Boyer and Barrett, 2005). Based on reliable covariance from empirical priors and inborn minimal mechanisms for differentiating foreground and conspecifics, the present state and future projections of the generative model facilitate (*via* actively inferred policies) the realization of preferred outcomes through the exploitation of local minima of uncertainty, i.e., generalized affordances. It is in this context that epistemic affordances play a key role and can be associated with intrinsic motivation, exploration, "motor babbling" and artificial curiosity in developmental neurorobotics (Schmidhuber, 2006, Baranes and Oudeyer, 2009). Put simply, being compelled to pursue FEM, uncertainty-reducing epistemically enriched policies ensure that agents quickly come to discover "what would happen if I did that."

Consider an example that works both literally and as a broad analogy to this generalized affordance process: the crossing of a roaring rapids *via* stepping stones. The rapids are in constant flux, but the fluctuations of the water also momentarily expose surface regions of the stones. In this sense, despite the high uncertainty brought about by the flux, the overlapping exposed surface regions for each stepping stone provide stable points—local minima of uncertainty. These local minima facilitate crossing the river, by which the preferred outcome of reaching the opposite bank is realized. In a literal sense, the stones are clearly conventional Gibsonian affordances, presented here as local minima of uncertainty in sequential states of the generative model. Analogically, the roaring rapids correspond to the general sensory flux of thermodynamic surface impingements, and the stepping stones correspond to any reliably invariant multimodal covariance established by empirical model updating. This sense of local minima also suggests a formal correspondence to the basins of attraction in neurodynamics (Freeman, 2012).

6.2. Skilled Expertise

By considering affordances in this light, we can demonstrate how affordance theory relates to arguments about skilled expertise from the perspective of phenomenological philosophy. The latter argues for the central role of embodiment as the basis of skilled expertise, in contrast to some conventional theories that view expertise in terms of a mastery of symbol systems and conditional rules (which, for historical or pragmatic reasons, can be commonly found in robotics/AI implementations). According to the most widely adopted embodied phenomenology theory of skill acquisition (Dreyfus and Dreyfus, 2005), there are five stages of progression from novice to expert, whether in, e.g., riding a bicycle, playing chess, or practicing medicine.

To briefly summarize these five stages, as the theory goes, a *novice* (in any domain) learns by appealing to basic rules that can indeed be expressed symbolically as propositions. Even with these conditional rules, the novice cannot necessarily discern what is relevant in the domain. This changes slightly in the next stage, when the *advanced beginner* continues to follow the rules, but gradually begins to notice what perceptions of the domain are relevant. Upon reaching the third stage, *competence*, the practitioner gains an appreciation of the vastness of domain-relevant nuances, along with the recognition that a list of rules could not be exhaustive; even if such a list could be near comprehensive, it would be too unwieldy to manage in real-time interaction.

Nevertheless, to cope with the domain, some rule-like responses remain helpful at this stage. The fourth stage, *proficiency*, finally overcomes the appeal to rule-like responses with an embodied ability to discern relevant situational nuance. However, the proficient practitioner continually reaches decision-making junctures that require a considered evaluation of different pathways forward. In the final stage, when *expertise* is attained, the expert seamlessly selects a pathway forward, rather than interrupting the "flow" (Csikszentmihalyi, 1990) for a considered evaluation. This form of embodied expertise is also described as "absorbed coping," referring to the phenomenological absorption in the interactive situation.

Without objecting to this characterization of embodied expertise as irreducible to symbols and rules, it is possible to explain the underpinnings of the stage progression using AIF simply by viewing the progression in reverse. If expertise is regarded as having a highly developed generative model of the agent/environment relationships within the domain, then the preferred future realized through active inference is the attainment of the implicit or explicit goal (cycling across the terrain or defeating the chess opponent). Through experience (i.e., empirical prior-based model updating of reliably invariant modality covariances), the agent discovers how to exploit the relevant affordances—the local minima of uncertainty in the generative model state transitions—to achieve the preferred outcome using domain-specific policies.[12]

By working backwards through the progression (moving from expert to novice), it becomes clear that without sufficient experience, the generative model has yet to become sufficiently "attuned" (a Gibsonian term) to the domain; some scaffolding is needed to stabilize the domain-specific interactions. The earlier the stage, the more scaffolding is needed, such that the novice relies almost exclusively on scaffolding (which need not be symbol and rule-based, as it could also be based on mimicry of experts). Any scaffolding presumably also serves to orient the non-expert practitioner to the relevant regularities that facilitate the progression. Note that, when learning to ride a bicycle, training wheels do not directly contribute to learning the cycling skill, but rather, they serve as supportive scaffolding to position the bicycle perpendicular to the ground until the relevant regularities for remaining perpendicular independently have been sufficiently learned.

An interesting robotics application of domain-specific sensorimotor skills is found in the notion of policy reuse and adaptation (Rosman et al., 2016). From an AIF perspective, this parallels an equivalent phenomenon in humans. For example, given the ability to ride a standard bicycle, and confronted with an unfamiliar old-fashioned penny-farthing, an agent could glean from the similar seat, handlebar, wheel, and pedal configuration that the bicycle-riding policy could be reused to ride the penny-farthing, with some necessary adjustments.

A real-world example in which a policy was adapted from a source to a particularly divergent target is the cultural advent of skateboarding, which was based on surfing.[13] Even though there are extreme differences between surfboard fins and skateboard wheels, ocean and pavement, the early skateboarders recognized the embodied motion similarities between the domains. In this case, a certain cross-domain policy identity is maintained through reuse and adaptation that focuses on the complex spatiotemporal nesting required in both practices involving body, board, and traversal surface: the interactive precision-weighting required for short timescale, rapid adjustments, and the simultaneous progressively longer timescales of extended maneuvering. The Gibsonian concept of "resonance" appears to be appropriately matched to such complex situated activity, in which the agent's multiscale embodied neurodynamics "resonate" with the multiscale environmental dynamics, following experiential attunement to the relevant regularities (Teques et al., 2017; cf. Raja, 2017).

6.3. The Frame Problem

At several points above, we have referred to the agent's identification of what is relevant or significant in a situation, which appears to run up against the frame problem. To recap, the frame problem holds that given actions that alter limited aspects of a situation, or given relevance-altering shifts in situational context, there is no clear mechanism to appeal to by which irrelevant situational aspects can be easily ignored. Dreyfus (1992) famously proposes that embodiment obviates the frame problem in a way that symbolic AI implementations cannot. He goes further still and proposes that even typical subsymbolic AI cannot overcome the problem; he finds some promise in Freeman's neurodynamics (Dreyfus, 2007), although his analysis of why this shows promise is limited. Given the convergences between Freeman's neurodynamics and AIF (Friston, 2008, 2010; De Ridder et al., 2014), it is not surprising that the latter should offer the robust response to the frame problem Dreyfus anticipated.

It is worth briefly restating the nature of neural computation in AIF, due to its substantial difference from the computation of input, symbols, propositional logic, and other common associations. Even the convenient shorthand used by neuroscientists and others that the brain "is" Bayesian or "implements" Bayesian models can lend itself to misunderstanding AIF's ontological commitments. Essentially, given synaptic connectivity and transmission patterns, it is possible to model them mathematically. It is rarely misunderstood when equations are used to descriptively model a planet's orbit in order to predict its positions—most people do not assume that this approach suggests the planet itself is computing anything (nor that the planet's material complexity is "reduced" or "eliminated" in the pragmatic abstraction of a mechanistic orbital model). Analogously, by appeal to the broader theoretical context of AIF, it can be stated that there are transformations in the dynamic neurobiological substrate in the service of the environmentally embedded body that can be

[12] In performing arts such as music, skilful policies may relate to actualized or simulated coordination in improvisation, performance, and compositional practices (see Linson, Forthcoming).

[13] The early skateboarders were "replicating on dry land the surfer's traverse across ocean surface and close sensing of changing wave forms. Through surf-related moves, skaters recombined body, board and terrain, simultaneously copying one activity (surfing) while initiating a second (skateboarding)" (Borden, 2001, pp. 31–33).

mathematically modeled in terms of probability distributions. Thus, embodied and embedded brain activity can be modeled as the computation of these distributions. That the calculations should be Bayes-approximate within AIF results from implicit pragmatic efficiency directives (arising from the constraints laid out in Sections 2 and 3), such as "extrapolate from experience" (empirical priors), "context matters" (hierarchical model architecture), and "when expectations are not met, re-assess" (respond to surprisal through model updating, precision-weighting, or abduction, depending on particulars about the accumulation of prediction error).

The frame problem, in its many incarnations, can be summarized in a single question: How does an agent know what is significant in an interactive situation? AIF answers with its own unique breakdown. The first level of the breakdown is that the agent can be either open or closed to potential significance. This is overlooked by most other accounts, which take openness to significance for granted, thereby missing the ecologically common phenomenon of habits. In AIF, habits can be regarded as context-free responses that are established by their invariance across multiple conditions (FitzGerald et al., 2014). When we act out of habit, we merely "go through the motions," suppressing any potential significance that might otherwise be contextually relevant.

Apart from habit, when the agent is open to potential significance, AIF points to a second-level breakdown of possible outcomes (when potential significance arises in a situation). Given that the active agent always entertains a repertoire of plausible policies within its generative model, there is a fundamental relationship between policy selection and the expected free energy within the policy or model space. Given that expected free energy scores the epistemic affordance of alternative policies on models, there is an inbuilt imperative to select *significant* or *relevant* actions. Significance in this instance is related to the epistemic, uncertainty-reducing component of expected free energy, while relevance can be construed in relation to prior preferences about ultimate actions. When a potentially significant aspect of the environment recruits a policy, it becomes relevant; this is equivalent to the notion of a "solicitation" in affordance theory and phenomenological philosophy (see Bruineberg and Rietveld, 2014; Bruineberg et al., 2016). In short, the significance or relevance is an integral aspect of FEM by which the frame problem is dissolved.

This argument rests upon appreciating that expected free energy can be decomposed into two parts (**Figure 1**). Variational free energy *per se* can always be decomposed into accuracy and complexity terms. This appeals to the Bayesian interpretation of variational free energy as an approximation to (or lower bound on) Bayesian model evidence. On this view, Bayesian model evidence is effectively *simplicity* plus *accuracy*.[14] But what about *expected* free energy? It transpires that *expected accuracy* is the expected probability of obtaining preferred outcomes, while *expected simplicity* is epistemic affordance, namely, the resolution of uncertainty or information gain afforded by the outcomes anticipated under any particular policy. This intrinsic value of a particular policy or model appears in many guises, most notably as intrinsic motivation in robotics (Oudeyer and Kaplan, 2007; Schmidhuber, 2010), the value of information in economics (Howard, 1966), and Bayesian surprise in models of exploration and visual searches (Schmidhuber, 1991; Itti and Baldi, 2009).

Ultimately, without the input/output model, the core difficulties associated with the frame problem—when to sample input, what to sample as input, what to do with input, or what becomes of fixed output—do not arise. There is only the generative model's accommodation of sensory perturbations in terms of hidden causes. By incorporating epistemic imperatives into the (Bayesian model) selection of policies in AIF, the broad frame problem never manifests. This is because novel information is not pre-screened for relevance, but instead is rendered relevant or significant when it leads to model updating or the selection of a new policy, and irrelevant or insignificant when it does neither. Note that the latter case holds irrespective of benefit or cost, given that the non-assimilation of novel information may be helpful (e.g., metabolic savings) or harmful (e.g., missed opportunity).

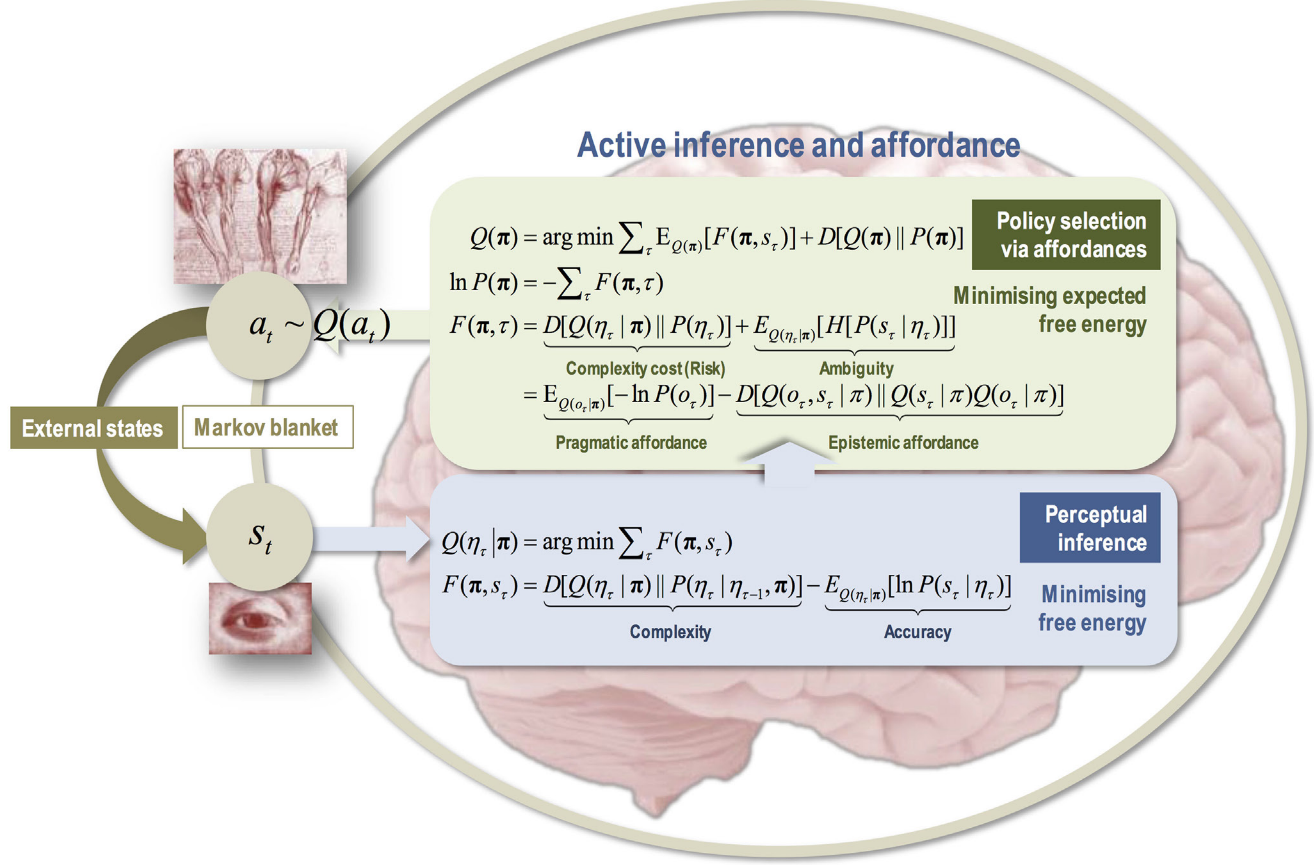
This approach also avoids concerns about the inadequacy of fixed representational encoding accounts of perception (Bickhard, 2008), given that in AIF, environmental information can serve multiple context-dependent relational roles in situated interaction (cf. Pylyshyn, 1999). Moreover, the logical frame problem is obviated by the probability distributions of the generative model—the agent interacts with the environment on the basis of expected model extrapolations, so continuous sensory sampling is unproblematic: samples either confirm expectations or produce surprisal (Mirza et al., 2016).

7. SELF-REFLECTIVE EPISTEMIC FORAGING: AN OPENING FOR CONSCIOUSNESS?

The reservoir of information present with respect to the self and the environment is inexhaustible. Only a small fraction is ever immediately relevant as adaptive behavioral control information. Thus, there are always new sources of potential relevance, as there are many possible signals in the noise (Dennett, 1991). While many discussions of AIF center on epistemic foraging in the environment, it is also possible to consider epistemic foraging of the self, also a rich source of signals in the noise (Seth, 2013; Seth and Friston, 2016).

Thus far, we have primarily addressed control information, noting that it can also be gleaned from conscious phenomenal sensation (Seth et al., 2012). Enhancing the generative model through exploration, also known as epistemic foraging, provides potential future control information. However, when new significance arises, it is not necessarily immediately subsumed as control information. Consider hearing a fellow diner's request to "pass the salt." Given situated language learning (Diessel, 2006), words

[14] Note that minimizing variational free energy implicitly minimizes complexity and associated computational costs—*via* Landauer's principle—that link thermodynamic free energy to variational free energy. In other words, the path of least variational free energy is, thermodynamically, Hamilton's path of least action.



$$Q(\boldsymbol{\pi}) = \arg\min \sum_\tau \mathrm{E}_{Q(\boldsymbol{\pi})}[F(\boldsymbol{\pi}, s_\tau)] + D[Q(\boldsymbol{\pi}) \,\|\, P(\boldsymbol{\pi})]$$

$$\ln P(\boldsymbol{\pi}) = -\sum_\tau F(\boldsymbol{\pi}, \tau)$$

$$F(\boldsymbol{\pi}, \tau) = \underbrace{D[Q(\eta_\tau \mid \boldsymbol{\pi}) \,\|\, P(\eta_\tau)]}_{\text{Complexity cost (Risk)}} + \underbrace{E_{Q(\eta_\tau \mid \boldsymbol{\pi})}[H[P(s_\tau \mid \eta_\tau)]]}_{\text{Ambiguity}}$$

$$= \underbrace{\mathrm{E}_{Q(o_\tau \mid \boldsymbol{\pi})}[-\ln P(o_\tau)]}_{\text{Pragmatic affordance}} - \underbrace{D[Q(o_\tau, s_\tau \mid \pi) \,\|\, Q(s_\tau \mid \pi) Q(o_\tau \mid \pi)]}_{\text{Epistemic affordance}}$$

$$Q(\eta_\tau \mid \boldsymbol{\pi}) = \arg\min \sum_\tau F(\boldsymbol{\pi}, s_\tau)$$

$$F(\boldsymbol{\pi}, s_\tau) = \underbrace{D[Q(\eta_\tau \mid \boldsymbol{\pi}) \,\|\, P(\eta_\tau \mid \eta_{\tau-1}, \boldsymbol{\pi})]}_{\text{Complexity}} - \underbrace{E_{Q(\eta_\tau \mid \boldsymbol{\pi})}[\ln P(s_\tau \mid \eta_\tau)]}_{\text{Accuracy}}$$

FIGURE 1 | Bayesian mechanics and active inference. This schematic summarizes the formal aspects of active inference in terms of minimizing variational free energy. It describes a generic (active) inference scheme that has been used in a wide variety of applications and simulations; ranging from games in behavioral economics (FitzGerald et al., 2015) and reinforcement learning (Schwartenbeck et al., 2015) through to language (Friston et al., 2017c) and scene construction (Mirza et al., 2016). The details of this scheme are not essential to understand the arguments in the main text: they are presented here for interested readers, with a special focus on how *affordance* emerges from minimizing (expected) free energy, under a generative model of the world. In this setup, discrete actions solicit a sensory outcome (*s*) that informs approximate posterior beliefs about external or hidden states of the world (η). This Bayesian belief updating can be described as minimizing variational free energy *F*(π, *s*) under a set of plausible policies (π). Here, a policy comprises a sequence of actions (*a*). The approximate posterior beliefs are then used to evaluate expected free energy *F*(π, τ) and subsequent beliefs about action; namely the *affordances* that underwrite policy selection. In other words, affordance corresponds to inference about action, where the most likely policy (to be selected) is the policy that minimizes expected free energy in the future. Q(η|π) denotes beliefs about hidden states in the future, given a particular policy and Q(π) are posterior beliefs about the policies currently being pursued. Free energy is just the difference between *complexity* and *accuracy*. In other words, an approximate posterior with a low free energy provides an accurate but simple explanation for sensory input. Expected free energy can be similarly decomposed into expected complexity (i.e., complexity cost or *risk*) and expected inaccuracy (i.e., *ambiguity*). Complexity can be regarded as the divergence (denoted by the Kullback–Leibler divergence *D*) between what one expects to happen under a particular policy and what one would prefer *a priori*. Ambiguity is the loss of a precise or definitive mapping between external states of the world and observed sensory states (as quantified by entropy, denoted by *H*). An alternative decomposition of expected free energy is in terms of *epistemic* and *pragmatic affordance*: see main text. Note a subtle but important aspect of this construction; namely, posterior beliefs about policies are based on their expected free energy, which includes the (path integral) of free energy *per se*. This is interesting from several perspectives. It means that the agent has to infer what it is doing and, implicitly, its own action. In other words, beliefs about action are distinct from the active states of the agent's Markov blanket (namely the sensory and action states that separate internal from external or hidden states). This means the agent has to predict how it will behave and then verify those predictions based on sensory evidence. This implicit inference means that the agent has to garner evidence for its own behavior. This is the role of the free energy. Namely, free energy *per se* provides evidence that a particular policy is being pursued, while expected free energy scores its prior probability. In summary, agents (will appear to) have beliefs about their behavior—beliefs that endow them with a sense of purpose, in virtue of the prior preferences that constitute risk. In effect, this enables agents to shape their sensorium. Please see Friston et al. (2017b) for technical details and Friston et al. (2017a) for a discussion of how this belief updating might be implemented in a brain.

provide evidence for the most apt generative model or policy (Lupyan and Clark, 2015), enhancing the control information for the relevant modification of the thermodynamic substrate (identifying, grabbing, and passing a nearby salt shaker). Nevertheless, the request is also appreciable as a phenomenal sensation that can be further epistemically foraged. For instance, the diner's shaky tone of voice might indicate an emotional state that was not immediately relevant to passing the salt, but may become relevant in social interaction, leading to an enquiry about their wellbeing (Filippi et al., 2017).

What should facilitate such inquiring? When time pressure is low, it is possible to reflectively evaluate information beyond its role in facilitating immediate adaptive behavior. AIF can describe this as the momentary decoupling of aspects of the model from the environment for self-reflective epistemic foraging, while potentially remaining partially environmentally engaged (e.g., thinking about the office during the commute). Having this ability would confer adaptive advantages, such as navigating complex social meaning, as well as more protracted forms of elaborate problem-solving (mentally revisiting a problem from different angles). This example also speaks to the trade-off between epistemic (expected simplicity) and pragmatic (expected accuracy) imperatives that underlie FEM in policy selection. In brief, the trade-off—not dissimilar to an exploration/exploitation trade-off—rests upon the precision of prior preferences. Generally, in a new situation, epistemic affordance would normally dominate policy selection until there is a comfortable familiarity with the lived context; prior preferences can then come into play. Crucially, these prior preferences are themselves inferred in deep (hierarchical) generative models.

A strong candidate for facilitating such self-reflection is also the most apparent correlate of self-consciousness: a mental buffer that underpins introspective awareness. This buffer can be regarded as the substrate of conscious mental simulation, imagination, and internal monolog. The latter would allow for forms of self-reflection, as well as the self-referential fine-tuning of adaptive behavior ("I must remain focused on the road!"). It is relatively uncontroversial to view simulation as contributing to adaptive behavior through mental rehearsal, and imagination as contributing to generating counterfactuals and exposing new affordances, while also enabling the suppression of conscious environmental coupling.

Whatever its genesis and other roles, consciousness appears to be crucial for epistemic foraging in the limitless source of signals in the noise of the self, in a manner wholly consistent with the information-bound AIF elaborated above. Note that bringing consciousness to the table presupposes a generative model of the future that necessarily entails a degree of selfhood and agency. This characteristic of generative models has been referred to as counterfactual richness or depth (Seth, 2015) to emphasize the deep and fictive nature of how (some) agents predict their world and behavior.

Moreover, from the AIF perspective, we can identify a feature that appears to be rare in the animal realm that could be plausibly robotically implemented. Our fundamentally thermodynamically constrained social origins imply a capacity for ethical considerations, at least concerning basic aspects of resource sharing (Cosmides et al., 2010). In this context, consciousness as a buffer for self-reflective epistemic foraging would underpin our ability to evaluate preferred outcomes and inferred policies from a space of possible state transitions—in other words, to evaluate ends and means to ends—on the basis of ethical considerations.

Through conscious, self-reflective epistemic foraging, a self-conscious agent can turn active inference inward, by nuancing model or policy selection to alter its current outcome preference. Also, when a preferred outcome has been selected, an agent can determine whether it ought to infer a policy alternative to the immediate, intuitively inferred policy it would have selected under time pressure.[15] (This can be thought of as the agent's self-referential policy to realize a preferred future in which *other* possible ends and means have been duly considered.) With the luxury of time, consciously aware self-reflective agents can individually and cooperatively aim for a deeply considered preferred future, to be reached *via* a deeply considered pathway.

The above speculations are indicative of the manner in which AIF can plausibly connect an agent's consciousness to its embedding in progressively larger social organizations. The mechanistic—yet radically non-reductive—explanatory underpinning of this embodied, embedded account of individuals and society inherently includes their openness to vast cultural proliferations and indeterminate futures.

8. CONCLUSION: AT THE CROSSROADS OF NATURAL AND ARTIFICIAL EMBODIED COGNITION

We have seen above why, in contrast to common assumptions, AIF *opposes* the mechanistic to the reductive. If AIF were applied to developing a humanoid robot that would approximate a human being, it is clear that its embodied apparatus must be more than just for show. The mechanical actuation would need to furnish the proprioceptive sensing aspect of the generative model that would exhibit reliably invariant covariance with exteroceptive sensing. For this extero- and proprioceptive coupling to be biomimetic, the sensing should have the same constraints as our biologically inherited apparatus, such as a limited visual range that is extended by bodily movement. Assuming a neuromorphic information integration apparatus were also implemented, we could expect robotic interoception to identify environmentally relevant quantities such as energy requirements ("hunger") and bodily damage ("pain").

So far, none of this would require consciousness, though it could achieve basic adaptive behavior. For a more deeply situated robot, we would need to add a minimal mechanism for distinguishing foreground from background, and one for differentiating between quasi-conspecifics (others of the same make or possibly humans as well). This could serve to fulfill the requirement of social grounding that would in principle pave the way for cooperative communication strategies, such as gesture and language.

With an appropriate buffer of interoceptive self-awareness, the robot could epistemically forage within this buffer for additional relevant signals than those it first identifies in the environment. Through the usual human routes of upbringing and education, it could also be taught to evaluate the consequences of its actions, to weigh preferred ends and available means by considering their potential impact on itself and others. The process of learning

[15] Time pressure is accommodated in active inference by appealing to Hamilton's principle of least action. In other words, it is the expected free energy over time that counts, where unexpected energy corresponds to an action. Put simply, for adaptive efficacy, it is much better to reduce free energy quickly, to an imperfect level, than to spend lots of time reducing it to its minimum.

to appreciate counterfactual outcomes would be enhanced by a capacity for valenced esthetic experiences ("emotions"). This suggests a broadly socially situated (humanlike) role for emotional regulation (see, e.g., Sell et al., 2017), which differs considerably from current robotic implementations of pseudo-emotional states (e.g., Moshkina et al., 2011).

It would be within reason to describe the set of processes in AIF as algorithms, which raises the question: what implications does this have for our understanding of humans? There have been many recent discussions of algorithmic bias in computer systems said to reflect the bias of the human system designers. This is not surprising, given any disembodied algorithm based on a reductive input/output model. With AIF, however, we can make sense of natural and artificial ecologically and socially situated embodied agents. Agents with this specification would interactively probe and learn the apparent regularities of their world. At the same time, with sufficient complexity, they would have the capacity to critically evaluate their own generalizations from past environmental exposure, to identify when forms of bias are detrimental, and to engage in meaningfully value-laden self-corrective recalibration (while of course this provides no guarantees, even for humans; see, e.g., Bang and Frith, 2017; Holroyd et al., 2017).

To summarize: by appeal to the principle of FEM, we can descriptively account for a long view that takes us from elementary particles to embodied biological agents. In an ecological context, the emergence and behavior of these agents—underpinned by a cybernetic relationship between thermodynamics and information—can be understood to plausibly facilitate the evolutionary development of life. On a long enough time scale, under contingent circumstances, FEM is sufficient to yield the coevolutionary development of mutually adaptive, highly complex agents and niches, as we see in human culture, especially in our pragmatic and epistemic foraging behavior, which fundamentally includes socially cooperative and self-reflective capacities. Taking all of this into account, AIF suggests a possible approach to the biomimetic modeling of human agents that in principle would exhibit humanlike embodied cognition. Such agents would plausibly be conscious in most senses of the word.

AUTHOR CONTRIBUTIONS

AL, AC, SR, and KF made substantial contributions to the conception and/or design of the work, its drafting and/or critical revision, and approved for publication its content.

ACKNOWLEDGMENTS

Parts of this material were presented in July 2017 at "Computation & Representation in Cognitive Science: Enactivism, Ecological Psychology & Cybernetics," held at the University of Sussex, organized by Jonny Lee, Joe Dewhurst, and Adrian Downey, and at "The World in Us: Gestalt Structure, Phenomenology and Embodied Cognition," held at the University of Edinburgh, organized by Dave Ward, Alistair Isaac, and Rosa Hardt. Many thanks to the organizers, participants, and speakers, especially Anthony Chemero, Ron Chrisley, and Julian Kiverstein, for their helpful discussions. Thanks as well to the participants at later presentations of this material to Karl Friston's Theoretical Neurobiology Group at the Wellcome Trust Centre for Neuroimaging, University College London, and to the department of General Systems Studies, University of Tokyo, with much appreciation to Takashi Ikegami for organizing the latter event. Many thanks to Jo Shaw for her advocacy of this work in its early stages during her directorship of the Institute for Advanced Studies in the Humanities (IASH), University of Edinburgh, and to current IASH director Steve Yearley, for the continued support during its completion. Additional thanks to David Carmel, Sam Cohn, and Steve Partridge for cross-disciplinary reinforcements. We also thank the editors and reviewers for their helpful feedback.

FUNDING

AL received initial funding for this research from the Institute for Advanced Studies in the Humanities at the University of Edinburgh, and open access publication funding from the University of Stirling. AC is funded in part by an ERC Advanced Grant, XSPECT (Ref: DLV-692739). SR is funded in part by the Xerox University Affairs Committee. KF is funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z).

REFERENCES

Adams, R. A., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218, 611–643. doi:10.1007/s00429-012-0475-5

Albrecht, S. V., Crandall, J. W., and Ramamoorthy, S. (2016). Belief and truth in hypothesised behaviours. *Artif. Intell.* 235, 63–94. doi:10.1016/j.artint.2016.02.004

Ashby, W. R. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica* 1, 83–89.

Ashourvan, A., Gu, S., Mattar, M. G., Vettel, J. M., and Bassett, D. S. (2017). The energy landscape underpinning module dynamics in the human brain connectome. *Neuroimage* 157, 364–380. doi:10.1016/j.neuroimage.2017.05.067

Auer, S., and Frenkel, D. (2001). Prediction of absolute crystal-nucleation rate in hard-sphere colloids. *Nature* 409, 1020–1023. doi:10.1038/35059035

Bang, D., and Frith, C. D. (2017). Making better decisions in groups. *R. Soc. Open Sci.* 4, 170193. doi:10.1098/rsos.170193

Baranes, A., and Oudeyer, P. Y. (2009). R-IAC: robust intrinsically motivated exploration and active learning. *IEEE Trans. Auton. Ment. Dev.* 1, 155–169. doi:10.1109/TAMD.2009.2037513

Barrett, H. C., Cosmides, L., and Tooby, J. (2010). Coevolution of cooperation, causal cognition and mindreading. *Commun. Integr. Biol.* 3, 522–524. doi:10.4161/cib.3.6.12604

Barto, A., Mirolli, M., and Baldassarre, G. (2013). Novelty or surprise? *Front. Psychol.* 4:907. doi:10.3389/fpsyg.2013.00907

Bechtel, W. (2014). "Cognitive biology: surprising model organisms for cognitive science," in *Proceedings of the Cognitive Science Society*, Vol. 36. Available at: http://www.escholarship.org/uc/item/0z82f8s6

Bechtel, W., and Abrahamsen, A. (2007). Explaining human freedom and dignity mechanistically: from receptive to active mechanisms. *J. Philos. Res.* 32, 43–66. doi:10.5840/jpr20073239

Bennett, C. H. (2003). Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Stud. Hist. Philos. Sci. B Stud. Hist. Philos. Mod. Phys.* 34, 501–510. doi:10.1016/S1355-2198(03)00039-X

Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour. *Br. J. Psychol.* 41, 68–80. doi:10.1111/j.2044-8295.1950.tb00262.x

Bickhard, M. H. (2008). Interactivism: a manifesto. *New Ideas Psychol.* 27, 85–95. doi:10.1016/j.newideapsych.2008.05.001

Bickhard, M. H., and Richie, D. M. (1983). *On the Nature of Representation: A Case Study of James Gibson's Theory of Perception*. New York: Praeger.

Borden, I. (2001). *Skateboarding, Space and the City: Architecture and the Body*. London: Bloomsbury Academic.

Boyer, P., and Barrett, H. C. (2005). "Domain specificity and intuitive ontology," in *The Handbook of Evolutionary Psychology*, ed. D. M. Buss (New York: Wiley), 96–118.

Brascamp, J. W., Knapen, T. H. J., Kanai, R., Noest, A. J., van Ee, R., and van den Berg, A. V. (2008). Multi-timescale perceptual history resolves visual ambiguity. *PLoS ONE* 3:e1497. doi:10.1371/journal.pone.0001497

Brooks, R. A. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.

Bruineberg, J., Kiverstein, J., and Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* 1–28. doi:10.1007/s11229-016-1239-1

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

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Front. Psychol.* 7:925. doi:10.3389/fpsyg.2016.00925

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

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

Chemero, A. (2003). An outline of a theory of affordances. *Ecol. Psychol.* 15, 181–195. doi:10.1207/S15326969ECO1502_5

Chemero, A. (2008). Self-organization, writ large. *Ecol. Psychol.* 20, 257–269. doi:10.1080/10407410802189372

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

Chemero, A., and Turvey, M. T. (2007). Gibsonian affordances for roboticists. *Adapt. Behav.* 15, 473–480. doi:10.1177/1059712307085098

Chen, J., Hasson, U., and Honey, C. J. (2015). Processing timescales as an organizing principle for primate cortex. *Neuron* 88, 244–246. doi:10.1016/j.neuron.2015.10.010

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. (2014). *(What Scientific Idea is Ready for Retirement?) The Input-Output Model of Perception and Action*. Edge.org. Available at: https://www.edge.org/response-detail/25394

Clark, A. (2015a). Radical predictive processing. *South. J. Philos.* 53, 3–27. doi:10.1111/sjp.12120

Clark, A. (2015b). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York: Oxford University Press.

Clark, A. (2016). Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs* 51, 727–753. doi:10.1111/nous.12140

Clark, A. (2017). "How to knit your own Markov blanket," in *Resisting the Second Law with Metamorphic Minds*, eds T. Metzinger and W. Wiese (Frankfurt am Main: MIND Group), 1–19.

Conant, R. C., and Ashby, W. R. (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

Cortina, M., and Liotti, G. (2010). The intersubjective and cooperative origins of consciousness: an evolutionary-developmental approach. *J. Am. Acad. Psychoanal. Dyn. Psychiatry* 38, 291–314. doi:10.1521/jaap.2010.38.2.291

Cosmides, L., Barrett, H. C., and Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9007–9014. doi:10.1073/pnas.0914623107

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Performance*. New York: Harper and Row.

de Haan, M., and Nelson, C. A. (1999). Brain activity differentiates face and object processing in 6-month-old infants. *Dev. Psychol.* 35, 1113–1121. doi:10.1037/0012-1649.35.4.1113

De Ridder, D., Vanneste, S., and Freeman, W. (2014). The Bayesian brain: phantom percepts resolve sensory uncertainty. *Neurosci. Biobehav. Rev.* 44, 4–15. doi:10.1016/j.neubiorev.2012.04.001

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron* 88, 2–19. doi:10.1016/j.neuron.2015.09.019

Dennett, D. C. (1991). Real patterns. *J. Philos.* 88, 27–51. doi:10.2307/2027085

Devaine, M., Hollard, G., and Daunizeau, J. (2014). Theory of mind: did evolution fool us? *PLoS ONE* 9:e87619. doi:10.1371/journal.pone.0087619

Diessel, H. (2006). Demonstratives, joint attention, and the emergence of grammar. *Cogn. Linguist.* 17, 463–489. doi:10.1515/COG.2006.015

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artif. Intell.* 171, 1137–1160. doi:10.1016/j.artint.2007.10.012

Dreyfus, H. L., and Dreyfus, S. E. (2005). Peripheral vision: expertise in real world contexts. *Org. Stud.* 26, 779–792. doi:10.1177/0170840605053102

Drossel, B., and Schwabl, F. (1992). Self-organized critical forest-fire model. *Phys. Rev. Lett.* 69, 1629–1632. doi:10.1103/PhysRevLett.69.1629

Dumas, G. (2011). Towards a two-body neuroscience. *Commun. Integr. Biol.* 4, 349–352. doi:10.4161/cib.4.3.15110

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

Evans, D. J., and Searles, D. J. (2002). The fluctuation theorem. *Adv. Phys.* 51, 1529–1585. doi:10.1080/00018730210155133

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., et al. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proc. R. Soc. B* 284, 20170990. doi:10.1098/rspb.2017.0990

FitzGerald, T. H., 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

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.* 8:457. doi:10.3389/fnhum.2014.00457

Fodor, J. A., and Pylyshyn, Z. W. (2002). "How direct is visual perception? Some reflections on Gibson's 'ecological approach,'" in *Vision and Mind: Selected Writings in the Philosophy of Perception*, eds A. Noë and E. Thompson (Cambridge, MA: MIT Press), 167–228.

Frank, S. A. (2012). Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J. Evol. Biol.* 25, 2377–2396. doi:10.1111/jeb.12010

Franz, A., and Triesch, J. (2010). A unified computational model of the development of object unity, object permanence, and occluded object trajectory perception. *Infant Behav. Dev.* 33, 635–653. doi:10.1016/j.infbeh.2010.07.018

Freeman, W. (2012). *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics*. Berlin: Springer Science & Business Media.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi:10.1371/journal.pcbi.1000211

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

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

Friston, K. (2011). What is optimal about motor control? *Neuron* 72, 488–498. doi:10.1016/j.neuron.2011.10.018

Friston, K. (2014). Active inference and agency. *Cogn. Neurosci.* 5, 119–121. doi:10.1080/17588928.2014.905517

Friston, K., and Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends Cogn. Sci.* 20, 500–511. doi:10.1016/j.tics.2016.05.001

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

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. J., Parr, T., and de Vries, B. (2017b). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi:10.1162/NETN_a_00018

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017c). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 77, 388–402. doi:10.1016/j.neubiorev.2017.04.009

Friston, K., Levin, M., Sengupta, B., and Pezzulo, G. (2015a). Knowing one's place: a free-energy approach to pattern regulation. *J. R. Soc. Interface* 12, 20141383. doi:10.1098/rsif.2014.1383

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

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

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biol. Cybern.* 102, 227–260. doi:10.1007/s00422-010-0364-z

Gibson, J. J. (1966). *The Senses Considered As Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979/1986). *The Ecological Approach to Visual Perception*. New York: Psychology Press.

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

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* 28, 2539–2550. doi:10.1523/JNEUROSCI.5487-07.2008

Hatfield, G. (2003). Representation and constraints: the inverse problem and the structure of visual space. *Acta Psychol.* 114, 355–378. doi:10.1016/j.actpsy.2003.07.003

Hohwy, J. (2013). *The Predictive Mind*. Oxford, New York: Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi:10.1111/nous.12062

Holroyd, J., Scaife, R., and Stafford, T. (2017). Responsibility for implicit bias. *Philos. Compass* 12, e12410. doi:10.1111/phc3.12410

Howard, R. A. (1966). Information value theory. *IEEE Trans. Syst. Sci. Cybern.* 2, 22–26. doi:10.1109/TSSC.1966.300074

Ito, H. C., and Ikegami, T. (2006). Food-web formation with recursive evolutionary branching. *J. Theor. Biol.* 238, 1–10. doi:10.1016/j.jtbi.2005.05.003

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.* 49, 1295–1306. doi:10.1016/j.visres.2008.09.007

Kang, Y. H. R., Petzschner, F. H., Wolpert, D. M., and Shadlen, M. N. (2017). Piercing of consciousness as a threshold-crossing operation. *Curr. Biol.* 27, 2285–2295.e6. doi:10.1016/j.cub.2017.06.047

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4:e1000209. doi:10.1371/journal.pcbi.1000209

Kirsh, D., and Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cogn. Sci.* 18, 513–549. doi:10.1207/s15516709cog1804_1

Kruschke, J. K. (2008). Bayesian approaches to associative learning: from passive to active learning. *Learn. Behav.* 36, 210–226. doi:10.3758/LB.36.3.210

Laland, K., Odling-Smee, J., and Endler, J. (2017). Niche construction, sources of selection and trait coevolution. *Interface Focus* 7, 1–9. doi:10.1098/rsfs.2016.0147

Lammert, H., Noel, J. K., and Onuchic, J. N. (2012). The dominant folding route minimizes backbone distortion in SH3. *PLoS Comput. Biol.* 8:e1002776. doi:10.1371/journal.pcbi.1002776

Leite, I., Martinho, C., and Paiva, A. (2013). Social robots for long-term interaction: a survey. *Int. J. Soc. Robot.* 5, 291–308. doi:10.1007/s12369-013-0178-y

Levine, R. D., and Tribus, M. (eds) (1978). *Maximum Entropy Formalism*, 1st Edn. 2nd Printing Edn. Cambridge, MA: The MIT Press.

Linson, A. (Forthcoming). "Moment's notice: models of time consciousness in philosophy and the cognitive sciences," in *Music and Consciousness*, Vol. 2, eds D. Clarke, E. Clarke, and R. Herbert (Oxford: Oxford University Press).

Lupyan, G., and Clark, A. (2015). Words and the world: predictive coding and the language-perception-cognition interface. *Curr. Dir. Psychol. Sci.* 24, 279–284. doi:10.1177/0963721415570732

Malamud, B. D., Morein, G., and Turcotte, D. L. (1998). Forest fires: an example of self-organized critical behavior. *Science* 281, 1840–1842. doi:10.1126/science.281.5384.1840

Mann, K., Gallen, C. L., and Clandinin, T. R. (2017). Whole-brain calcium imaging reveals an intrinsic functional network in *Drosophila*. *Curr. Biol.* 27, 2389–2396.e4. doi:10.1016/j.cub.2017.06.076

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Henry Holt and Company.

Maturana, H. R. (1995). "Biology of self-consciousness," in *Consciousness: Distinction and Reflection*, ed. G. Tratteur (Naples: Bibliopolis), 145–175.

Maxwell, J. C. (1871). *Theory of Heat*. London: Longmans, Green, and Co.

McCarthy, J., and Hayes, P. J. (1969). "Some philosophical problems from the standpoint of artificial intelligence," in *Machine Intelligence 4*, eds B. Meltzer and D. Michie (Edinburgh University Press), 463–502.

Minsky, M. (1974). *A Framework for Representing Knowledge*. Cambridge, MA: MIT.

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

Modayil, J., and Kuipers, B. (2008). The initial development of object knowledge by a learning robot. *Rob. Auton. Syst.* 56, 879–890. doi:10.1016/j.robot.2008.08.004

Modayil, J., White, A., and Sutton, R. S. (2014). Multi-timescale nexting in a reinforcement learning robot. *Adapt. Behav.* 22, 146–160. doi:10.1177/1059712313511648

Moshkina, L., Park, S., Arkin, R. C., Lee, J. K., and Jung, H. (2011). TAME: time-varying affective response for humanoid robots. *Int. J. Soc. Robot.* 3, 207–221. doi:10.1007/s12369-011-0090-2

Moulin, C., and Souchay, C. (2015). An active inference and epistemic value view of metacognition. *Cogn. Neurosci.* 6, 221–222. doi:10.1080/17588928.2015.1051015

Newell, A., Shaw, J. C., and Simon, H. A. (1959). "Report on a general problem-solving program," in *Proceedings of the International Conference on Information Processing* (Paris), 256–264.

Orlandi, N. (2017). Bayesian perception is ecological perception. *Philos. Top.* 44, 327–351. doi:10.5840/philtopics201644226

Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front Neurorobot.* 1:6. doi:10.3389/neuro.12.006.2007

Pezzulo, G. (2017). "Tracing the roots of cognition in predictive processing," in *Philosophy and Predictive Processing*, eds T. K. Metzinger and W. Wiese (Frankfurt am Main: MIND Group), 1–20.

Pezzulo, G., and Cisek, P. (2016). Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends Cogn. Sci.* 20, 414–424. doi:10.1016/j.tics.2016.03.013

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

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–190. doi:10.1017/S0140525X04000056

Pierce, D., and Kuipers, B. J. (1997). Map learning with uninterpreted sensors and effectors. *Artif. Intell.* 92, 169–227. doi:10.1016/S0004-3702(96)00051-3

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

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav. Brain Sci.* 22, 341–365. doi:10.1017/S0140525X99002022

Raja, V. (2017). A theory of resonance: towards an ecological cognitive architecture. *Minds Mach.* 72, 1–23. doi:10.1007/s11023-017-9431-8

Ramamoorthy, S., Salamon, A. Z., and Santhanam, R. (2012). Macroscopes: models for collective decision making. *arXiv:1204.3860 [cs]*. Collective Intelligence 2012: Proceedings. Available at: http://arxiv.org/abs/1204.3860

Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T., and Reissland, N. (2017). The human fetus preferentially engages with face-like visual stimuli. *Curr. Biol.* 27, 1825–1828.e3. doi:10.1016/j.cub.2017.05.044

Rosman, B., Hawasly, M., and Ramamoorthy, S. (2016). Bayesian policy reuse. *Mach. Learn.* 104, 99–127. doi:10.1007/s10994-016-5547-y

Sahin, E., Çakmak, M., Dogar, M. R., Ugur, E., and Ucoluk, 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

Schmidhuber, J. (1991). "Curious model-building control systems," in *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, Vol. 2 (Singapore, Singapore), 1458–1463.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect. Sci.* 18, 173–187. doi:10.1080/09540090600768658

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi:10.1109/TAMD.2010.2056368

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* 4:710. doi:10.3389/fpsyg.2013.00710

Schwartenbeck, P., FitzGerald, T., Mathys, C., Dolan, R., Kronbichler, M., and Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Sci. Rep.* 5, 16575. doi:10.1038/srep16575

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., et al. (2017). The grammar of anger: mapping the computational architecture of a recalibrational emotion. *Cognition* 168, 110–128. doi:10.1016/j.cognition.2017.06.002

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013). Information and efficiency in the nervous system—a synthesis. *PLoS Comput. Biol.* 9:e1003157. doi:10.1371/journal.pcbi.1003157

Sergent, C., Wyart, V., Babo-Rebelo, M., Cohen, L., Naccache, L., and Tallon-Baudry, C. (2013). Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Curr. Biol.* 23, 150–155. doi:10.1016/j.cub.2012.11.047

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. (2015). "Inference to the best prediction," in *Open MIND*, eds T. K. Metzinger and J. M. Windt (Frankfurt am Main: MIND Group), 1–8.

Seth, A. K., and Edelman, G. M. (2004). Environment and behavior influence the complexity of evolved neural networks. *Adapt. Behav.* 12, 5–20. doi:10.1177/105971230401200103

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371, 20160007. doi:10.1098/rstb.2016.0007

Seth, A. K., Izhikevich, E., Reeke, G. N., and Edelman, G. M. (2006). Theories and measures of consciousness: an extended framework. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10799–10804. doi:10.1073/pnas.0604347103

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Front. Psychol.* 2:395. doi:10.3389/fpsyg.2011.00395

Shagrir, O. (2010). Marr on computational-level theories. *Philos. Sci.* 77, 477–500. doi:10.1086/656005

Shanahan, M. (1997). *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge, MA: MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x

Simon, H. A. (1957). *Administrative Behavior: A Study of Administrative Processes in Administrative Organization*. New York: MacMillan.

Sliwa, J., and Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science* 356, 745–749. doi:10.1126/science.aam6383

Southgate, V., Csibra, G., Kaufman, J., and Johnson, M. H. (2008). Distinct processing of objects and faces in the infant brain. *J. Cogn. Neurosci.* 20, 741–749. doi:10.1162/jocn.2008.20052

Sprevak, M. (2005). "The frame problem and the treatment of prediction," in *Computing, Philosophy and Cognition*, eds L. Magnani and R. Dossena (London: King's College Publications), 349–359.

Stahl, A. E., and Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science* 348, 91–94. doi:10.1126/science.aaa3799

Still, S., and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory Biosci.* 131, 139–148. doi:10.1007/s12064-011-0142-z

Street, S. E., Navarrete, A. F., Reader, S. M., and Laland, K. N. (2017). Coevolution of cultural intelligence, extended life history, sociality, and brain size in primates. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7908–7914. doi:10.1073/pnas.1620734114

Teques, P., Araújo, D., Seifert, L., del Campo, V. L., and Davids, K. (2017). The resonant system: linking brain–body–environment in sport performance☆. *Prog. Brain Res.* 234, 33–52. doi:10.1016/bs.pbr.2017.06.001

Ugur, E., Oztop, E., and Sahin, E. (2011). Goal emulation and planning in perceptual space using learned affordances. *Rob. Auton. Syst.* 59, 580–595. doi:10.1016/j.robot.2011.04.005

Ugur, E., Şahin, E., and Oztop, E. (2012). "Self-discovery of motor primitives and learning grasp affordances," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vilamoura: IEEE), 3260–3267.

Ullman, S. (1980). Against direct perception. *Behav. Brain Sci.* 3, 373–381. doi:10.1017/S0140525X0000546X

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *Int. J. Philos. Stud.* 16, 323–349. doi:10.1080/09672550802113235

Yaeger, L. S. (2009). How evolution guides complexity. *HFSP J.* 3, 328–339. doi:10.2976/1.3233712

Yaeger, L. S. (2013). Identifying neural network topologies that foster dynamical complexity. *Adv. Complex Syst.* 16, 1350032. doi:10.1142/S021952591350032X

Yaeger, L. S., and Sporns, O. (2006). "Evolution of neural structure and complexity in a computational ecology," in *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems* (Cambridge, MA: MIT Press/Bradford Books), 330–336.

Yahiro, T., Kataoka, N., Nakamura, Y., and Nakamura, K. (2017). The lateral parabrachial nucleus, but not the thalamus, mediates thermosensory pathways for behavioural thermoregulation. *Sci. Rep.* 7, 5031. doi:10.1038/s41598-017-05327-8

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., et al. (2017). Same story, different story: the neural representation of interpretive frameworks. *Psychol. Sci.* 28, 307–319. doi:10.1177/0956797616682029

Yoshida, W., Dolan, R. J., and Friston, K. J. (2008). Game theory of mind. *PLoS Comput. Biol.* 4:e1000254. doi:10.1371/journal.pcbi.1000254

Conflict of Interest Statement: 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.

APPENDIX

Glossary of Terms

In Bayesian statistics and machine learning, several common terms have technical meanings. This glossary defines the way in which we use key terms in the current article.

Free-energy: an information theory measure that bounds (is greater than) the surprise on sampling some data, given a generative model.

Entropy: the average surprise of outcomes sampled from a probability distribution or density. A density with low entropy means, on average, the outcome is relatively predictable. High entropy denotes unpredictability and uncertainty.

Surprise, *surprisal*, or *self-information*: the negative log-probability of an outcome. An improbable outcome is, therefore, surprising. Negative surprise is the same as *log evidence*; namely, the logarithm of Bayesian model evidence.

Bayesian surprise: a measure of salience based on the divergence between the posterior and prior probability densities. It measures the information gain obtained by updating the priors to posteriors.

[Kullback–Leibler] Divergence: information divergence, information gain, or relative entropy. The divergence is a (non-commutative) measure of the difference between two probability distributions.

Generative model: a probabilistic model that generates consequences (i.e., data) from their causes (i.e., model parameters). A generative model is also known as a forward model and is usually specified in terms of the likelihood of getting some data given their causes (parameters of a model) and priors on the parameters.

Prior: the probability distribution or density over the causes of data that encode beliefs about those causes prior to observing the data.

Empirical prior: priors that are induced by hierarchical models; they provide constraints on the recognition density is the usual way but depend on the data.

Conditional density or *posterior density*: the probability distribution over causes or model parameters, given some data; i.e., a probabilistic mapping from observed consequences to causes. In Bayesian inference, the prior is updated—on the basis of observations—to become a posterior, according to Bayes rule.

Model evidence: in Bayesian statistics, the model evidence is the probability that observed data were generated by a particular generative model. The negative logarithm of model evidence is surprise or self-information in information theory.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 12 March 2018
doi: 10.3389/frobt.2018.00020

Self in NARS, an AGI System

Pei Wang, Xiang Li and Patrick Hammer*

Department of Computer and Information Sciences, Temple University, Philadelphia, PA, United States

This article describes and discusses the self-related mechanisms of a general-purpose intelligent system, NARS. This system is designed to be adaptive and to work with insufficient knowledge and resources. The system's various cognitive functions are uniformly carried out by a central reasoning-learning process following a "non-axiomatic" logic. This logic captures the regularities of human empirical reasoning, where all beliefs are revisable according to evidence, and the meaning of concepts are grounded in the system's experience. NARS perceives its internal environment basically in the same way as how it perceives its external environment although the sensors involved are completely different. Consequently, its self-knowledge is mostly acquired and constructive, while being incomplete and subjective. Similarly, self-control in NARS is realized using mental operations, which supplement and adjust the automatic inference control routine. It is argued that a general-purpose intelligent system needs the notion of a "self," and the related knowledge and functions are developed gradually according to the system's experience. Such a mechanism has been implemented in NARS in a preliminary form.

Keywords: general intelligence, non-axiomatic logic, self-awareness, self-control, self-organization, consciousness

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Bas Steunebrink,
NNAISENSE SA, Switzerland
Haris Dindo,
Università degli Studi di Palermo, Italy

****Correspondence:***
Pei Wang
pei.wang@temple.edu

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *31 October 2017*
Accepted: *16 February 2018*
Published: *12 March 2018*

Citation:
Wang P, Li X and Hammer P (2018) Self in NARS, an AGI System. Front. Robot. AI 5:20. doi: 10.3389/frobt.2018.00020

1. INTRODUCTION

Phenomena and functions like "self-awareness," "self-control," "self-reference," and "self-consciousness" are closely related to human intelligence, cognition, and thinking, and the related topics have been discussed in various fields (Hofstadter, 1979; Blackmore, 2004).

In the study of artificial intelligence (AI), although these issues have been addressed by the pioneers (Simon, 1962; Minsky, 1985; McCarthy, 1995), they nevertheless have been rarely considered in the technical works, as shown by the lack of coverage of these topics in the common textbooks (Luger, 2008; Russell and Norvig, 2010; Poole and Mackworth, 2017). The difficulty of realizing these functions in a machine is both technical and theoretical, as there is no widely accepted theory about them, and even their definitions are highly controversial.

On the contrary, researchers in the emerging field of artificial general intelligence (AGI) widely consider these functions as necessary for general intelligence and have proposed various ways to cover hem in AGI systems (Schmidhuber, 2007; Baars and Franklin, 2009; Bach, 2009; Shapiro and Bona, 2010; Chella and Manzotti, 2012; Thórisson, 2012; Goertzel, 2014; Rosenbloom et al., 2016). As these approaches are based on very different considerations and typically tangled with the other functions in the system, it is hard to compare them to say which one is the best.

The focus of this article is the relevant aspects of NARS (non-axiomatic reasoning system), a formal model of general intelligence, which has been mostly implemented and is under testing and tuning. In the following, the conceptual design of NARS is introduced first, then the parts mostly relevant to "self" are described in more detail. Finally, the major design decisions are compared with the related works.

2. NARS OVERVIEW

NARS (non-axiomatic reasoning system) is an AGI-designed framework of a reasoning system. The project has been described in many publications, including two books (Wang, 2006, 2013), so it is only briefly summarized here.

2.1. Theoretical and Strategic Assumptions

The working definition of "intelligence" in NARS is different from that in mainstream AI, where "Intelligence" is usually taken as an ability to solve problems that are only solvable by the human brain. A computer agent can obtain this ability by developing domain-specific solutions. Instead, NARS is designed according to the belief that "Intelligence" is *the ability for a system to adapt to its environment and to work with insufficient knowledge and resources*. It requires the system to have the capacities of accepting unanticipated problems and events, making real-time responses, working with finite resources, and learning from its experience in an application domain.

The behaviors of NARS are based on past experiences and generated by interacting with the environment in real time; therefore, the solutions provided by the system to the problems are usually not the optimum solutions but the best solution that the system can find at the moment. The system could always do better with more resources and knowledge, especially in a relatively stable environment. Compared to the other theories of rationality, the most significant feature of this "relative rationality" is the *Assumption of Insufficient Knowledge and Resources*, hereafter **AIKR**. Concretely, the following three features are demanded by AIKR, with respect to the problems to be solved by the system:

- **Finite:** The system is able to work with constant information-processing capacity, in terms of processor speed, storage space, etc.
- **Real time:** The system is able to deal with problems that show up at any moment and the utility of their solutions may decrease over time.
- **Open:** The system is able to accept input data and problems of any content, as long as they are expressed in a format recognizable by the system.

Due to the time and resources restriction, and also the uncertainty about the coming problems, NARS usually cannot consider all possibilities when facing a problem, but will only consider some important and relevant possibilities, judged according to the system's experience.

According to AIKR, NARS does not treat the storage space of itself as infinite. The mechanism of forgetting is a special feature of NARS to deal with limited storage space. Some beliefs or tasks will be removed from the storage of NARS when their *priority* (to be introduced later) is below a threshold.

AIKR is a fundamental assumption abstracted from the study of human problem solving in the real world. Humans obtain knowledge by learning and summarizing past experience. When humans deal with problems that they do not know how to solve at the moment, they will attempt to solve them with the help of relevant knowledge. *This ability is exactly what we consider intelligence, which is characterized not by what problems it can solve, but the restriction under which the problems are solved.*

The research goal of NARS is to design and build a computer system that can adapt to its environment and solve problems under AIKR. This is different from the objectives of the mainstream AI projects, which are specific problem-solving abilities. The aim of NARS is to build a system with a given learning ability (at the meta-level) that allows the system to acquire various problem-solving skills from its experience.

Although being a reasoning system is neither a necessary nor sufficient condition for being intelligent, a reasoning system can provide a suitable framework for the study of intelligence, as it forces the system to be general purpose, instead of being domain specific. Reasoning is at a more abstract level than other low-level cognitive activities, and it is obviously a critical cognitive skill that qualitatively distinguishes human beings from other animals.

Many cognitive processes such as planning, learning, decision-making, etc., can be formulated as types of reasoning; therefore, an intelligent system designed in the framework of a reasoning system can be extended to cover them easily. As a reasoning system follows a logic, each step of processing must be justifiable independently. As a result, inference steps can be linked at run time in novel orders to handle novel problems. This is a major reason why NARS is designed as a reasoning system.

2.2. Knowledge Representation

As a reasoning system, NARS uses a formal language called "Narsese" for knowledge representation, which is defined by a formal grammar given in the study by Wang (2013). To fully specify and explain this language is beyond the scope of this article, so in the following, only the directly relevant part is introduced informally and described briefly.

The logic used in NARS belongs to a tradition of logic called "term logic," where the smallest component of the representation language is a "term," and the simplest statement has a "subject-copula-predicate" format, where the subject and the predicate are both terms.

The basic form of statement in Narsese is *inheritance statement*, which has a format "$S \rightarrow P$," where S is the subject term, and P is the predicate term, the "→" is the *inheritance* copula, which is defined as a reflexive and transitive relation from one term to another term. The intuitive meaning of "$S \rightarrow P$" is "S is a special case of P" and "P is a general case of S." For example, statement "*robin* → *bird*" intuitively means "Robin is a type of bird."

We define the *extension* of a given term T to contain all of its known special cases and its *intension* to contain all of its known general cases. Therefore, "$S \rightarrow P$" is equivalent to "S is included in the extension of P," and "P is included in the intension of S."

The simplest, or "atomic," form of a term is a *word*, that is, a string of characters from a finite alphabet. In this article, typical terms are common English nouns like *bird* an *animal*, or mixed by English letters, digits 0–9, and a few special signs, such as hyphen("-") and underscore ("_"), but the system can also use other alphabets or use terms that are meaningless to human beings, such as "drib" and "aminal."

Beside atomic terms, Narsese also includes *compound terms* of various types. A compound term (*con*, C_1, C_2, …, C_n) is formed by a term connector, *con*, and one or more component terms (C_1, C_2, …, C_n). The term connector is a logical constant with predefined meaning in the system. Major types of compound terms in Narsese include the following:

- **Sets:** Term {*Tom*, *Jerry*} is an *extensional set* specified by enumerating its instances; term [*small*, *yellow*] is an *intensional set* specified by enumerating its properties.
- **Intersections and differences:** Term (*bird* ∩ *swimmer*) represents "birds that can swim"; term (*bird* – *swimmer*) represents "birds that cannot swim."
- **Products and images:** The relation "John is the uncle of Zack" is represented as "({*John*} × {*Zack*}) → *uncle-of*," "{*John*}→ (*uncle-of* / ◊ {*Zack*})," and "{*Zack*}→ (*uncle-of* / {*John*} ◊)," equivalently.[1] Here, ◊ is a placeholder, which indicates the position in the *uncle-of* relation the subject term belongs to.
- **Statement:** "John knows soccer balls are round" can be represented as a *higher-order statement* "{*John*}→ (*know* / ◊ {*soccer-ball* → [*round*]})," where the statement "*soccer-ball* → *[round]*" is used as a term.
- **Compound statements:** Statements can be combined using term connectors for disjunction("∨"), conjunction("∧"), and negation("¬"), which are intuitively similar to those in propositional logic, but not defined using truth-tables.[2]

Several term connectors can be extended to take more than two component terms, and the connector is often written before the components rather between them, such as (× {*John*}{*Zack*}).

Beside the *inheritance* copula ("→", "is a type of"), Narsese also has three other basic copulas: *similarity* ("↔", "is similar to"), *implication* ("⇒", "if-then"), and *equivalence* ("⇔", "if-and-only-if"), and the last two are used between statements.

In NARS, an *event* is a statement with temporal attributes. Based on their occurrence order, two events E_1 and E_2 may have one of the following basic temporal relations:

- E_1 happens before E_2
- E_1 happens after E_2
- E_1 happens when E_2 happen

More complicated temporal relations can be expressed by taking about the subevents of the events.

Temporal statements are formed by combining the above basic temporal relations with the logical relations indicated by the term connectors and copulas. For example, implication statement "$E_1 \Rightarrow E_2$" has three temporal versions, corresponding to the above three temporal orders, respectively[3]:

- E_1 /⇒ E_2
- E_1 \⇒ E_2
- E_1 |⇒ E_2

All the previous statements can be seen as Narsese describing things or events from a third-person view. Narsese can also describe the actions of the system *itself* with a special kind of event called *operation*. An operation is an event directly realizable by the system itself *via* executing the associated code or command.

Formally, an operation is an application of an operator on a list of arguments, written as $op(a_1, \ldots, a_n)$ where *op* is the operator, and $a_1, \ldots, a_n$ is a list of arguments. Such an operation is interpreted logically as statement"(× {*SELF*} {a_1} … {a_n}) → *op*," where *SELF* is a special term indicating the system itself, and *op* is an operator that has a procedural interpretation. For instance, if we want to describe an event "The system is holding key_001," the statement can be expressed as "(× {*SELF*} {*key*_001})→ *hold*."

Overall, there are three types of sentences defined in Narsese:

- A **judgment** is a statement with a truth value and represents a piece of new knowledge that system needs to learn or consider. For example, "*robin* → *bird* ⟨*f, c*⟩," where the truth value ⟨*f, c*⟩ will be introduced in the next section.
- A **question:** is a statement without a truth value, and represents a question to be answered according to the system's beliefs. For example, if the system has a belief "*robin* → *bird*" (with a truth value), it can be used to answer question "*robin* → *bird*?" by reporting the truth value, as well as to answer the question "*robin* → ?" by reporting the truth value together with the term *bird*, as it is in the intension of *robin*. Similarly, the same belief can also be used to answer question "? → *bird*" by reporting the truth value together with the term *robin*.
- A **goal** is statement without a truth value, and represents a statement to be realized by executing some operations, according to the system's beliefs. For example, "(× {*SELF*} {*door*_001}) → *open*!" means the system has the goal to open the *door*_001 or to make sure that *door*_001 is opened. Each statement of goal always associates with a "desire-value," indicating the extent to which the system hopes for a situation where the statement is true.

The *experience* of NARS consists of a stream of input sentences of the above types.

2.3. Experience-Grounded Semantics

When studying a language, semantics relates the items in the language to the environment in which the language is used. It answers questions like "What is the meaning of this term?," or "What is the truth value of that statement?"

Since NARS is designed under AIKR, the truth value of a statement measures its extent of evidential support, rather than that of agreement with a corresponding fact. NARS does not determine the truthfulness of its knowledge with respect to a static and completely described environment. Since the environment changes over time, there is no guarantee that the past is always identical to the future. Hence, in NARS, the truth of each statement and the meaning of each term are grounded on nothing but the system's experience. The formal definition of this

[1] This treatment is similar to the set-theoretic definition of "relation" as set of tuples, where it is also possible to define what is related to a given element in the relation as a set. For detailed discussions, see the studies by Wang (2006, 2013).

[2] The definitions of disjunction and conjunction in propositional logic do not require the components to be related in content, which lead to various issues under AIKR. In NARS, such a compound is formed only when the components are related semantically, temporally, or spatially. See the study by Wang (2013) for details.

[3] Here, the direction of the arrowhead is the direction of the implication relation, while the direction of the slash is the direction of the temporal order. In principle, copulas like "/⇐" can also be defined, but they will be redundant. For more discussion on this topic, see the study by Wang (2013).

semantics and discussions of its implications can be found in the studies by Wang (2005, 2013) and are only briefly summarized in the following.

As mentioned previously, in Narsese, "*robin* → *bird*" states that "Robin is a type of bird," and it is equivalent to saying that the extension of *robin* is included in the extension of *bird*, as well as the intension of *bird* is included in the intension of *robin*. Therefore, if a term is in the extension (or intension) of both *robin* and *bird*, then its existence supports the statement or provides positive evidence. On the contrary, if a term is in the extension of *robin* but not the extension of *bird*, or is in the intension of *bird* but not the intension of *robin*, it provides negative evidence for the statement.

For a given statement, we use w^+, w^-, and w to represent the amount of positive, negative, and total evidence, respectively. Based on them, a two-dimensional truth value is defined as a pair of real numbers $\langle f, c\rangle$ for the measurements. Here, f is called the *frequency* of the statement and is defined as the proportion of positive evidence among total evidence, that is, $f = w^+/w$. The value c is called the *confidence* of the statement and is defined as the proportion of current evidence among total amount of evidence at a moment in the future after new evidence of a certain amount is collected, that is, $c = w/(w + k)$, where $k \geq 1$. This constant k is a "personality parameter" and is explained further in the study by Wang (2013). The value of k can be seen as a unit of evidence that decides how fast the c value increases as new evidence comes, and in the following, we use the default $k = 1$ to simplify the discussion. Roughly speaking, *frequency* represents the uncertainty of the statement, and *confidence* represents the uncertainty of the frequency (Wang, 2001). Defined in this way, *truth value* in NARS is "experience-grounded."

Similarly, the *meaning* of a term is defined as its extension and intension, so it is determined by how it is related to other terms in the system's experience. As the experience of a system grows over time, the truth value of statements and the meaning of terms in the system change accordingly. This *experience-grounded semantics* (EGS) is fundamentally different from the traditional *model-theoretic semantics*, since it defines *truth value* and *meaning* according to a (dynamic and system-specific) experience, rather than a (static and system-independent) model. In the simplest implementation of NARS, its *experience* is a stream of Narsese sentences, which will be summarized to become the system's *beliefs*, which is also called the system's *knowledge*. This semantics is formally defined and fully discussed in the study by Wang (2005, 2006).

2.4. Inference Rules

The logic followed by NARS is NAL (non-axiomatic logic), and its inference rules use Narsese sentences as premises and conclusions. A recent version of NAL is formalized and justified in the study by Wang (2013). What is described in the following is only a small part of NAL that is directly related to the current topic.

NAL uses formal inference rules to recursively derive new knowledge from existing knowledge, which consists of statements with truth values, indicating the experienced relations between terms and the strength of these relations. Each inference rule has a truth value function that calculates the truth value of the conclusion according to the evidence provided by the premises.

In terms of the type of reasoning, inference rules of NARS are divided into three categories:

- **Local rules:** These rules do not derive new statements. Instead, the conclusion comes out from a revision or selection of the premises.
- **Forward rules:** New judgments are produced from a given judgment and a relevant belief.
- **Backward rules:** New questions (or goals) are produced from a given question (or goal) and a relevant belief.

In the following, these three groups of rules are introduced in that order.

Under AIKR, NARS may have inconsistent beliefs, that is, the same statement may obtain different truth values according to different evidential bases. When the system locates such an inconsistency, it either uses the *revision* rule (if the evidence bases are disjoint) or the *choice* rule (if the evidence bases are not disjoint). The revision rule accepts two judgments about the same statement as premises and generates a new judgment for the statement, with a truth value obtained by pooling the evidence of the premises. Consequently, the frequency of the conclusion is a weighted sum of those premises, and the confidence is higher than those of the premises. The choice rule simply choose the premise that has more positive evidence and less negative evidence, while preferring simpler candidates.[4]

As a term logic, typical inference rules in NAL are *syllogistic*, and each rule takes two premises (with one common term) to derive a conclusion (between the other two terms). The NAL rules of this type include *deduction*, *induction*, and *abduction*, similar to how the three are specified by Peirce (1931), although the truth value of every statement is extended from {0, 1} to [0,1] × (0,1). These three inference rules are the most basic forward rules of NAL, where *M*, *P*, and *S* represent arbitrary terms:

Deduction	Induction	Abduction
$M \rightarrow P \langle f_1, c_1\rangle$ $S \rightarrow M \langle f_2, c_2\rangle$	$M \rightarrow P \langle f_1, c_1\rangle$ $M \rightarrow S \langle f_2, c_2\rangle$	$P \rightarrow M \langle f_1, c_1\rangle$ $S \rightarrow M \langle f_2, c_2\rangle$
$S \rightarrow P \langle f, c\rangle$	$S \rightarrow P \langle f, c\rangle$	$S \rightarrow P \langle f, c\rangle$

Different forward inference rules have different truth value functions that calculate $\langle f, c\rangle$ from $\langle f_1, c_1\rangle$ and $\langle f_2, c_2\rangle$. These functions are established in the study by Wang (2013), and here, we do not describe the actual functions, but merely divide the inference rules into two groups, according to the maximum *confidence* value of the conclusions:

- **Strong inference:** The upper bound of confidence is 1. Among the rules introduced so far, only the *deduction* rule belongs to this group.

[4] The truth value function of the choice rule and the syntactic complexity of a term is defined in the study by Wang (2013).

- **Weak inference:** The upper bound of the confidence is $1/(1 + k) \leq 1/2$. The *abduction* and *induction* rules belong to this group.

The *weak inference* rules in NARS usually carry out *learning*, where each piece of evidence generates a weak conclusion, and strong conclusions are accumulated by the *revision* rule from many weak conclusions. This is why "learning" and "reasoning" are basically the same process in NARS (Wang and Li, 2016).

NAL has other syllogistic rules and also has *compositional* rules to build compound terms to capture the observed patterns in experience. For example, from "*swan* → *bird* $\langle f_1, c_1 \rangle$" and "*swan* → *swimmer* $\langle f_2, c_2 \rangle$," a rule can produce "*swan* → (∩, *bird, swimmer*) $\langle f, c \rangle$."

The inference rules of NAL can be used in both *forward inference* (from existing beliefs to derived beliefs) and *backward inference* (from existing beliefs and questions/goals to derived questions/goals). For each forward inference rule that from two judgments J_1 and J_2 to derive a conclusion J, a backward inference rule can be established that takes J_1 and a question on J as input and derives a question on J_2, because an answer for the derived question can be used together with J_1 to provide an answer to the original question. For example, if the question is "*robin* → *animal*?," and there is a related belief "*robin* → *bird* $\langle f, c \rangle$," then a derived question "*bird* → *animal*?" can be generated. The backward inference on goals is similar.

2.5. Inference Control

Equipped with the inference rules of NAL, NARS can carry out the following types of inference tasks:

- To absorb new experience into the system's beliefs, as well as to spontaneously derive some of their implications.
- To answer the input questions and the derived questions according to the system's beliefs.
- To achieve the input goals and the derived goals by executing the related operations according to the system's beliefs.

Under AIKR, new tasks can enter the system at any time, each with its own time requirement, and its content can be any Narsese sentence. Working in such a situation, usually NARS cannot perfectly accomplish all tasks in time, but has to allocate its limited time and space resources among them and to dynamically adjust the allocation according to the change of context, the feedback to its actions, and other relevant factors.

In the memory of NARS, beliefs and tasks are organized into *concepts*, according to the terms appearing in them. Roughly speaking, for a term T, concept C_T refers to all beliefs and tasks containing T. For example, the beliefs on "*robin* → *bird*" are referred to within concepts C_{robin} and C_{bird}, as well as other relevant concepts. A "concept" in NARS is a unit of both storage and processing and models the concepts found in human thinking.

To indicate the relative importance of concepts, tasks, and beliefs to the system, *priority* distributions are maintained among them. The priority of an item (concept, task, or belief) summarizes the attributes to be considered in resource allocation, including its intrinsic quality, usefulness in history, relevance to the current context, etc. Consequently, items with higher priority values will get more resources.

Bag is a data structure specially designed for resource allocation in NARS. A certain type of data items is contained in a bag with a constant capacity, with a priority distribution among the items maintained. There are three basic operations defined in a bag:

- *put(item)*: put an item into the bag, and if the bag is full, remove an item with the lowest priority
- *get(key)*: take an item from the bag with a given key that uniquely identifies the item
- *select()*: select an item from the bag, and the probability for each item to be selected is positively correlated with its priority value

NARS works by repeating an inference cycle consisting of the following major steps:

1. Select a concept within the memory
2. Select a task referred by the concept
3. Select a belief referred by the concept
4. Derive new tasks from the selected task and belief by the applicable inference rules
5. Adjust the priority of the selected belief, task, and concept according to the context and feedback
6. Selectively put the new tasks into the corresponding concepts and report some of them to the user

All selections in the above steps are probabilistic, and the probability for an item to be selected is positively correlated to its priority value. Consequently, the tasks will be processed in a time-sharing manner, with different speeds. For a specific task, its processing does not follow a predetermined algorithm, but it is the result of many inference steps, whose combination is formed at run time, so is usually neither predictable nor repeatable accurately, as both the external environment and the internal state of the system change in a non-circular manner.

3. "SELF" IN NARS

In this section, we focus on the aspects of NARS that are directly relevant to self-awareness and self-control.

3.1. Self in Various Forms

"Self" takes multiple forms in NARS. Some of the relevant properties are addressed by different mechanisms built into the system, and some others are shown in the system's learning process, including the following:

- **Higher-order statements**: As described previously, the higher-order statements in Narsese cover "statement about statement," "knowledge about operations," etc., which are often taken as functions of "metacognition" (Cox, 2005). Since such knowledge is typically about individual statements or operations and not about the system as a whole, and they are not the focus of this article. This type of knowledge usually is processed using inference rules analogical to these used on the statement level. For more details, see the studies by Wang (2006, 2013).

- **Intrinsic mechanisms:** As a part of the inference control process, NARS constantly compares the certainty of beliefs and dynamically allocates its resources among competing tasks. Even though the relevant mechanisms are indeed at a meta-level with respect to beliefs and tasks, they are implicitly embedded in the code, so not generally accessible to the system's deliberation nor can they be modified by the system itself. Therefore, they describe a constant aspect of the system itself that is not reflected in the object-level beliefs of the system.
- **Experience-grounded semantics:** As mentioned previously, the system's beliefs and concepts are built from the viewpoint of the system itself rather than as an objective model of the world. In this sense, all beliefs in NARS are subjective, and all concepts have idiosyncratic meanings to various extents. Consequently, the system's behaviors can be explained and predicted only when the unique experience of the system itself is taken into consideration.

Although the above mechanisms are all related to the system itself in a broad sense, they nevertheless can be described without explicitly using the notion of "self." In the following, the discussion will focus on "self" in a narrow sense, where a reference to the system as a whole becomes necessary.

3.2. The "Self"-Concept

NARS' beliefs about itself start at its built-in operations. As mentioned above, operation $op(a_1, \ldots, a_n)$ corresponds to a relation that the system can establish between itself and the arguments, so it is equivalent to statement "$(\times\{SELF\}\ \{a_1\}\ldots\{a_n\}) \rightarrow op$" (where the subject term is a *product* term written in the prefix format), since it specifies a relation among the arguments plus the system identified by the special term *SELF*.

Similar to the case of logic programming (Kowalski, 1979), here the idea is to uniformly represent declarative knowledge and procedural knowledge. So in NARS, knowledge about the system itself is unified with knowledge about others. For instance, the operation "open this door" is represented as "$(\times\{SELF\}\{door\_1\}) \rightarrow open$," so the inheritance copula encodes that the relation between {*SELF*} and {*door_1*} is a special case of opening. On the other hand, "John opened this door" is represented as "$(\times\{John\}\{door\_1\}) \rightarrow open$" (tense omitted to simplify the discussion). In this way, imitation can be carried out by analogical inference.

According to experience-grounded semantics (EGS), in NARS, the meaning of a concept is gradually acquired from the system's experience. However, EGS does not exclude the existence of innate concepts, beliefs, and tasks. In the above example, *SELF* is such a concept, with built-in operations that can be directly executed from the very beginning of the system's life. Such operations depend on the hardware/software of the host system, so are not specified as parts of NARS, except that they must obey the format requirements of Narsese. According to EGS, in the initial state of NARS, the meaning of a built-in operation is procedurally expressed in the corresponding routine, while the meaning of *SELF* consists of these operations. To the system, "*I* am whatever I can do and feel," since in NARS *sensation* (converting signals into terms) and *perception* (organizing terms into compounds) are also carried out by operations.

As the system begins to have experience, the meaning of every concept will be more or less adjusted as it is experienced, directly or indirectly. For a built-in operation, the system will gradually learn its preconditions and consequences, so as to associate it with the goals it can achieve and the context where it can be used. It is like we learn how to raise our hand first and then know it as a way to get the teacher's attention. The *SELF*-concept will be enriched in this way, as well as through its relations with other concepts representing objects and other systems in the outside environment.

Therefore, self starts from "what I can do and feel" to include "what I am composed of," "how I look like," "what my position is in the society," etc. The notion "self" does not have a constant meaning determined by a denotation or definition. Instead, the system gradually learns who it is, and its self-image does not necessarily converge to a "true self." Since the change of meaning of a concept is done *via* the additions, deletions, and revisions of its relations with other concepts, the system's identity (determined by all the relations) is relatively stable in a short period, although in its whole life the system may change greatly, even to the extent of unrecognizable when compared to a previous image of itself. Under AIKR, the system is open to all kinds of experience, so in the design of NARS, there is no restriction on the extent of these changes.

When NARS is used to serve a practical purpose, we often need to bind its behaviors, but it should be achieved *via* the control of the system's experience, rather than by designing the system in a special way, as also described in the study by Bieger and Thórisson (2016).

3.3. Mental Operations

An operation may be completely executed by the actuator of the host system (e.g., a NARS-controlled robot raises a hand or moves forward) or partly by another coupled system or device (e.g., a NARS-controlled robot pushes a button or issues a command to another system). NARS has an interface for such "external" operations to be registered. Consequently, all kinds of operations to be used in a "plug-and-play" manner, i.e., to be connected to the system at run time by a user or the system itself. A learning phase is usually needed for an operation to be used properly and effectively, as NARS will gradually learn its preconditions and consequences.

In principle, operations are not necessarily demanded in every NARS implementation, except a special type of "mental" operations that operate on the system's own "mind." There are several groups of mental operations in the current design, including

- **Task generation:** An inference task in NARS can either be input or derived recursively from an input task. The derivation process does not change the type of the task (judgment, question, or goal). However, in certain situations, a task needs to be generated from another one of a different type. For example, a new judgment ("It is cold.") may trigger a new goal ("Close the window!"). This relation is represented as an implication statement where the consequent is not a statement, but an operation call, similar to a production rule (Luger, 2008).

- **Evidence disqualification:** By default, the amount of evidence for every belief accumulates over time. Therefore, although the frequency value of the belief may either increase or decrease (depending on whether the new evidence is positive or negative), its confidence value increases monotonically. This treatment is supplemented by a mental operation that allows the system to doubt a belief of itself by decreasing its confidence value to a certain extent.
- **Concept activation:** The resource allocation mechanism of NARS already implements a process similar to activation spreading in neural networks (Russell and Norvig, 2010). When a new task is added into a concept, the priority of the concept is increased temporarily, and inference in the concept may cause derived tasks to be sent to its neighbors, so their priority levels will be increased, too. As a supplement, a mental operation allows the system to pay attention to a concept without new tasks added, so as to allow the system to deliberately consider a concept.

In general, mental operations supplement and influence the automatic control mechanism, and let certain actions be taken as the consequence of inference. Mental operations contribute to the system's self-concept by telling the system what is going on in its mind and allow the system to control its own thinking process to a certain extent. For instance, the system can explicitly plan its processing of a certain type of task. After the design and implementation phases, the system needs to learn how to properly use its mental operations, just like it needs to learn about the other (external) operations.

3.4. Internal Experience

In NARS, "experience" refers to the system's input streams. In the simplest implementation of NARS, the system has only one input channel, where the experience from the channel is a stream of the form S_1, T_1, S_2, T_2, …, S_n, T_n, where each S_i is a Narsese sentence, with T_i to be the time interval between it and the next sentence. A buffer of a constant size n holds the most recent experience.

In more complicated implementations, there are also "sensory" channels, each accepting a stream of Narsese terms from a sensory organ. Here, a sensor can recognize a certain type of signal, either from the outside of the system (such as visual or audio signals) or from the inside of the system, either from its body (somatosensory) or from its mind (mental). An internal channel provides a certain type of "internal experience." Somatosensory input will be especially important for a robotic system, as it needs to be aware of its energy level, network connection status, damages in parts, etc.

A mental sensation may come from the execution of a mental operation. Also, there are mental sensations appearing as the traces of the system's inference activity. During each inference cycle, the system "senses" the concept that was selected for processing, as well as the derivation relationship between tasks. Later, this experience can be used to answer questions such as "What has been pondered?" or "Where does that conclusion come from?," asked either by the system itself or by someone else. This information can also be used in future inference activities.

On the input buffers, the system carries out certain perceptive reasoning to form compound terms corresponding to the spatiotemporal patterns of the input. There is also a global buffer that holds a stream of Narsese sentences that integrate inputs from all the channels. In this aspect, the external and internal experiences are handled basically in the same manner.

A special type of belief formed in perception is the temporal implications between the mental events sensed within the system and the outside events observed by the system. The system will believe that it is some of its ideas that "cause" a certain action to be performed in its environment, and such beliefs will coordinate its "mind" and its "body." This is also arguably the origin of the notion of "causation" within the system. For a detailed discussion on temporal and causal inference in NARS, see the study by Wang and Hammer (2015).

The internal experience of NARS is the major source of its self-knowledge. Under AIKR, this type of knowledge is also uncertain and incomplete and is under constant revision. Furthermore, it is subjective and from the first-person perspective. In these aspects, NARS is fundamentally different from the "logical AI" approach toward self-knowledge, where the system is assumed as "having certain kinds of facts about its own mental processes and state of mind" (McCarthy, 1995).

3.5. Feeling and Emotion

According to AIKR, NARS needs to deal with different tasks with limited time and other resources. To ask the designer to provide a general optimizing algorithm to manage resources for all the possible situations is obviously impossible, and this is one of the reasons why NARS needs a mechanism to learn how to manage its resources and to make quick responses in various circumstances, all by itself. In the human mind, emotion and feeling play major roles in situation appraisal and behavior control, which are also desired in computer systems (Arbib and Fellous, 2004). In NARS, we have built a preliminary mechanism to carry out similar functions.

NARS has a basic satisfaction–evaluation mechanism at the event level. Every event has a truth value and a desire value, expressing the current status and what the system wants it to be, respectively. The closeness between them is called "satisfaction," which indicates a basic appraisal of an individual aspect of the situation. The value of "satisfaction" is in the range [0, 1], where 0 means "completely unsatisfied," 1 for "completely satisfied," and the other cases are in between.

Also there is system-level satisfaction, as the accumulation of recent event-level satisfactions, which represents an appraisal of the overall situation. Technically, this value is evaluated in every working cycle by adjusting the overall satisfaction value using the satisfaction value of the event just processed. This system-level satisfaction indicates the system's extent of "happiness" or "pleasure," and it plays multiple roles within the system, such as influencing the resource allocation.

To make the system aware of the values of these satisfaction indicators, some "feeling" operators are implemented, which reflect these satisfaction values into the internal experience of the system, so as to involve them explicitly into the inference processes. This happens by the usage of reserved terms and statements, which form the category of "emotional concepts" within the memory of the system. These emotional concepts provide a perception of emotions within NARS to the system itself, just like how the perceptive concepts summarize the system's experience when interacting with the outside world.

These emotional concepts interact with other concepts as generic (unemotional) concepts would, leading to the generation of compounds by the inference process, be represented by concepts that combine the emotional aspect with other aspects of the situation. Being unsatisfied about an event may be caused by other systems or the system itself, may be about the past or the future, may be controllable or inevitable, etc., and all these differences will lead to different categorization about the situation. For example, simply speaking, *regret* is the combination of negative emotion (unsatisfied situation) with other concepts like "things happened in the past" and "things caused by my own behaviors." You will not feel *regret* about bad things that might happen in the future or caused by the behaviors of someone else.

In addition, desire value is extended to non-event concepts according to their correlation with overall satisfaction. For example, an object will be liked by the system if the appearing of this object consistently concurs with high satisfaction level, and the contrary ones will be "disliked" by the system. Of course, there are many other things for which the system has little emotion. These different attitudes mainly come from the system's experience and will influence the system's treatment to the concepts.

In summary, in NARS, emotional information appears in two distinct forms:

- At the "subconscious level," it appears as desire values and satisfaction values. They are outside of the experience of the system, since these values do not form statements the system could reason about.
- At the "conscious level," it appears as events expressed using emotional concepts. They are inside of the experience of the system, since they are represented as statements that are considered in the inference process of the system.

Emotional information in both forms contributes to the system's internal processes, as well as to the system's external behaviors.

The emotional concepts in experience are processed as other concepts in inference. Consequently, they categorize the objects and situations according to the system's appraisal and allow the system to behave accordingly. For instance, the system may develop behavior patterns for "danger," even though each concrete danger has very different sources and causes.

The "emotion-specific" treatments also happen at the subconscious level, where the emotional information is used in various processes.

- The desire values of concepts are taken into account in attention allocation, where the concepts associated with strong feeling (extreme desire values) get more resources than those with weak feeling (neutral desire values). These desire values not only help the system to judge how long data items should be stored in memory but also how much priority they should be given when under consideration.
- After an inference step, if a goal is relatively satisfied, its priority is decreased accordingly and the belief used in the step gets a higher priority because of its usefulness. This way, already satisfied goals get less attention by the system, while relevant knowledge that satisfied these goals tends to be kept in memory longer, with the related concepts "liked" by the system.
- In decisions made, the threshold on confidence is lower in high emotional situations to allow quick responses. This is especially desired in situations where there is no lot of time available to react.
- The overall satisfaction is used as a feedback to adjust the priority values of data items (concepts, tasks, beliefs), so that the ones associated with positive feeling are rewarded, and the ones associated with negative feeling punished. In this way, the system shows a "pleasure-seeking" tendency, and its extent can be adjusted by a system parameter. This pleasure-seeking tendency can be considered as a motivation that is not directly based on any task, but as a "meta-task."
- When the system is relatively satisfied, it is more likely to create new goals, while when the system is unhappy about the current situation, it is more likely to focus on the existing goals that have not been achieved.

Overall, the system's feelings and emotions consist of a major part of its internal experience and contribute to its self-control. Emotion also plays roles in communication and socialization, but they, as well as topics like the self-control of emotion, are beyond the scope of this article.

3.6. Examples

Here, we illustrate a few examples using the Open-NARS[5] implementation of NARS. To simplify the description, the examples are slightly edited to remove the attributes not discussed in this article (such as the tense of the sentences), and before each Narsese sentence, the type of the sentence and a rough English translation are added. The ASCII symbols in the actual input/output are not the same as the logical symbols in the publications (including the above sections), but since their correspondence is hinted by their similarity and suggested by the English translation, the format will not be explained in detail, except the following:

- Judgments, questions, and goals in Narsese end with ".", "?", and "!", respectively.
- Prefix "^" indicates an operator, prefix "#" indicates an anonymous term, and prefix "$" indicates a variable term that can be substituted by another term.
- When the truth value of an input judgment or the desire value of an input goal is unspecified, the default $\langle 1, 0.9\rangle$ is used.

The first example demonstrates learning from observing the actions of another agent. Let's assume that Michael sells a car and that it is observed that he is rich after that. Later, when the system

[5] Source code, working examples, and documentations of Open-NARS can be found at http://opennars.github.io/opennars/.

gets the goal "to be rich," it will want to sell a car, too, as it guesses that whatever worked for Michael will also work for itself.

```
Input: "Michael sells a car."
<(*,{Michael},car) --> ^sell>.
Input: "Michael gets rich."
<{Michael} --> [rich]>.
Derived: "After someone sells a car, one gets rich."
<(&/,<(*,$1,car) --> ^sell>) =/> <$1 --> [rich]>>.
%1.00;0.31%
Input: "I want to be rich!" <{SELF} --> [rich]>!
Derived: "I want to sell a car!"
<(*,{SELF},car) --> ^sell>! %1.00;0.28%
```

This example shows that the system uses a temporal relation as evidence for a causal relation, which of course often leads to mistakes. In NARS, such mistakes are corrected by further negative evidences, that is, when the system learns other car-selling events that do not bring richness to the seller. This is also how the system resolves competing explanations and predictions, that is, by accumulating evidence on the competing hypotheses and choosing the best supported one.

The next example illustrates how the system summarizes its experience in relation to itself. In particular, it shows that picking up trash together with the knowledge that itself is a robot leads to the formation of a compound concept that contributes to the meaning of itself as a "robot that picks up trash":

```
Input: "I am a robot."
<{SELF} --> robot>.
Input: "I pick up trash."
<(*,{SELF},trash) --> ^pick>.
Derived: "I am somebody who picks up trash."
<{SELF} --> (/,^pick,_,trash)>.
Input: "What two things characterize you?"
<{SELF} --> (&,?1,?2)>?
Answer: "That I am a robot who picks up trash."
<{SELF} --> (&,(/,^pick,_,trash),robot)>. %1.00;0.81%
```

The intermediate result that transformed the second input statement into an *inheritance* statement about itself was crucial here. The same happens with mental operations. The case where the system wonders about whether cats are animals illustrates that:

```
Input: "I wonder whether cats are animals."
<(*,{SELF},<cat --> animal>) --> ^wonder>.
Input: "What am I?"
<{SELF} --> ?1>?
Answer: "I am somebody who wonders whether cats
are animals."
<{SELF} --> (/,^wonder,_,<cat --> animal>)>.
%1.00;0.90%
```

Such a wondering event is part of the internal experience of the system and is generated by a question:

```
<cat --> animal>?
```

For this to happen, the question task needs to exceed a certain priority value, meaning the system has to consider it as sufficiently important to the current situation.

The next examples show other motivational and emotional aspects of the system, such as the usage of a "*feel*" operator. Besides that, it shows the system's capability to consider the related event in question answering:

```
Input: "I don't want to get hurt."
(--,<{SELF} --> [hurt]>)!
Input: "When running away from a close wolf, I won't
get hurt."
<(&/,<(*,{SELF}, wolf) --> close_to>,
<(*,{SELF}) --> ^run>) =/> (--,<{SELF} --> [hurt]>)>.
Input: "I am close to wolf_1 now."
<(*,{SELF}, {wolf_1}) --> close_to>.
Input: "Wolf_1 is a wolf"
<{wolf1} --> wolf>.
Execution: "I run away."
<(*,{SELF}) --> ^run>!
Input: "I did not get hurt."
(--,<{SELF} --> [hurt]>).
```

The system deriving that running away from the wolf is satisfying:

```
"Feel the amount of satisfaction!"
(^feelSatisfied,{SELF})!
Feedback: "I am relatively satisfied."
<{SELF} --> [satisfied]>. %0.65;0.90%
Input: "How can I be satisfied?"
<?how =/> <{SELF} --> [satisfied]>>?
Answer: "Running away when a wolf is close makes me
satisfied."
<(&/,<wolf --> (/,close_to,{SELF},_)>,<(*,{SELF})
--> ^run>)
    =/> <{SELF} --> [satisfied]>>. %0.59;0.40%
```

Such satisfaction-related events can lead to emotion-based decisions, and, as the example shows, compound term can be composed by combining these events with other knowledge in the system.

In animals, there is usually an innate link between getting hurt by another animal and experiencing fear by future appearances of this kind of animal. Also, the response to fear, namely to run away, is usually an innate reaction and at the same time a successful strategy to survive. The following example demonstrates this case:

```
Innate belief: "If you are close to something that
frightens you, run away"
<(&/,<(*,{SELF}, #1) --> close_to>,<(*,#1,{SELF})
--> frightens>)
=/> <(*,{SELF},<(*,{SELF}) --> ^run>) --> ^want>>.
Innate belief: "If something hurts you, it frightens
you."
<<(*,$1,{SELF}) --> hurt> =/> <(*,$1,{SELF}) -->
frightens>>.
Innate belief: "If something frightens you, you feel
fear.
<<(*,#1,{SELF}) --> frightens> =|> <(*,{SELF},fear)
--> feel>>.
Input: "You are getting hurt by a wolf."
<(*,wolf,{SELF}) --> hurt>.
```

From here, it is expected that the system learned to be fearful of wolves and that it runs away whenever it encounters one.

```
Input: "You are close to a wolf."
<(*,{SELF}, wolf) --> close_to>.
Input: "How do you feel?"
<(*,{SELF},?what) --> feel>?
Answer: "I feel fear."
<(*,{SELF},fear) --> feel>. %1.00;0.29%
Execution: "I run away."
<{SELF} --> ^run>!
```

Given this encoding, the system can also be asked what frightens it:

```
Input: "What frightens you?"
<(*,?1,{SELF}) --> frightens>?
Answer: "The wolf frightens me."
<(*,wolf,{SELF}) --> frightens>. %1.00;0.43%
```

4. COMPARISONS AND DISCUSSIONS

In this section, the design decisions in NARS that are directly related to "self" are explained and compared with the alternatives.

4.1. The Need for a Self

Are self-awareness and self-control really required in an intelligent system? Why are such functions absent in most of the AI systems developed so far?

Like many controversies in AI, the different opinions on this matter can be traced back to the different understandings of "AI" (Wang, 2008). As the mainstream AI aims at the solving of specific problems, the systems are usually equipped with problem-specific algorithms, which embed knowledge about the problem domain, but not about the system itself, as the properties of the problem solver are usually irrelevant to the problem-solving process.

Even in learning systems that do not demand algorithms to be manually coded, they are still approximated by generalizing training data (Flach, 2012). In general, such systems have little need to add itself into the picture, as the solutions should only depend on the data to be learned, not the learner. Even metacognition can be carried out without an explicit "self"-concept involved (Cox, 2005)—when all the decisions are made by the system, it is unnecessary to explicitly state that.

In AGI systems, the situation is different. Here, we have projects aimed at simulating the human mind according to psychological theories, such as LIDA (Franklin, 2007) and MicroPsi (Bach, 2009), which surely need to simulate the self-related cognitive functions, simply because the well-known roles they play in human cognition (Blackmore, 2004).

In the function-oriented AGI projects, self-awareness and self-control are introduced to meet the requirements for the system, rather than sorely to be human-like. For instances, GLAIR is able to "represent and reason about beliefs about itself" (Shapiro and Bona, 2010). Sigma has the function of "architectural self-monitoring" (Rosenbloom et al., 2016). When facing varying problems, an AGI has to know itself and be able to adjust itself, so as to meet the changing situations. Since the existing AGI systems have very different overall designs, the exact form of the self-related functions differ greatly, and it is hard to compare and judge them in details without taking the whole system into consideration.

In general, NARS is more similar to GLAIR and Sigma than to LIDA and MicroPsi, as it is designed to realize a certain understanding of intelligence, which is generalized away from its realization in human beings. For NARS, the need for self-awareness and self-control follows from its working definition of intelligence, that is, adaptation under AIKR (Wang, 2008). To adapt to the environment and to carry out its tasks, the system needs to know what it can do and how it is related to the objects and other systems in the environment, and an explicitly expressed *SELF*-concept organizes all the related tasks and beliefs together, so as to facilitate reasoning and decision-making.

It may be argued that there is already a "self" in many AI systems, as knowledge in the system is often conceptually "about itself," "by itself," or "for itself." Why bother to explicitly spell that out and to separate it from other knowledge?

It is indeed the case that many AI systems have self-knowledge without explicitly talking about itself, but taking it as the default. For example, many works under the name of "metacognition" (Cox, 2005) have knowledge about various algorithms within the system itself and use this knowledge to select a proper one for the current problem. Although this process is self-reflective by nature, the systems typically does not have an explicitly represented "self." Instead, the processes are separated into "object-level" and "meta-level," where the latter monitor and control the former (Cox, 2005; Marshall, 2006).

Although an "implicit self" is enough for many problems, an explicitly represented *self*-concept provides many advantages desired in general-purpose AI that must adapt to various situations. This idea is not really new, as it can be at least traced back to McCarthy (1995), who promoted the idea of "making robots conscious of their mental states." In NARS, the *SELF*-concept provides a flexible unit for the representation and processing of self-knowledge coming from various sources and in different forms, although it does not cover all the self-related functions. As a reasoning system, this design allows NARS to uniformly represent and process knowledge about the system itself and about the other systems. As shown by the previous examples, imitation can be directly carried out as analogical reasoning, by substituting another system by *SELF*.

4.2. Self-Awareness

The self-knowledge of NARS shares many features as the system's knowledge about the outside environment.

All types of knowledge in NARS are organized into concepts. According to the semantics of NARS, the meaning of a concept (or a term naming a concept) is normally determined by its relation with other concepts (or terms). While for most concepts such relations are all acquired from the system's experience, the system is not necessarily born with a blank memory. Each built-in operation contributes meaning to the concept of *SELF*, by relating the system as a whole to the events it can perceive and/or realize. Starting from these operations, the *SELF*-concept will eventually involve beliefs about

- what the system can sense and do, not only using the built-in operations, but also the compound operations recursively composed from them, as well as the preconditions and consequences of these operations;
- what the system desires and actively pursues, that is, its motivational and emotional structure;
- how the system is related to the objects and events in the environment, in terms of their significance and affordance to the system;
- how the system is related to the other systems, that is, the "social roles" played by the system, as well as the conversions in communication and interactions.

All these aspects will make the system's self-concept richer and richer, even to the level of complexity that we can meaningfully talk about its "personality," that is, what makes this system different from the others, due to its unique nature and nurture. It is possible to measure the complexity of a concept in terms of its conceptual relations whose truth value is stable (high *confidence*) and unambiguous (extreme *frequency*), although such a measurement does not mean much, as the intuitive richness of a concept also depends on many other factors, such as the quality and diversity of the concepts it relates to, and so on.

This treatment is fundamentally different from identifying "self" with a physical body or a constant mechanism within the system. The spatial scope of self is mainly determined by the range of the system's sensors and effectors, which can distribute in distinct locations.

According to our approach, "self" is not left completely to a mysterious "emergent process," neither. In NARS, the concept *SELF* starts with a built-in core, then evolves according to the system's experience. In the process, the self-concept organizes the relevant beliefs and tasks together to facilitate self-awareness and self-control. This is consistent with Piaget's theory that a child learns about self and environment by coordinating sensing (such as vision and hearing) with actions (such as grasping, sucking, and stepping) and gradually progresses from reflexive, instinctual action at birth to symbolic mental operations (Piaget, 1963).

NARS treats *SELF* like other concepts in the system, except that it is a "reserved word," which has innate associations with the built-in operations, including the mental operations. NARS also treats internal and external experience uniformly, so self-awareness and self-control are nothing magical or mysterious, but are similar to how the system perceives and acts upon the external environment.

An important type of self-knowledge is provided by the emotion and feeling mechanism of NARS. As mentioned previously and described in detail in the study by Wang et al. (2016), such a mechanism is introduced into NARS, not for giving the system a "human face," but for appraising the current situation and dealing with it efficiently.

McCarthy (1995) concluded that "Human-like emotional structures are possible but unnecessary for useful intelligent behavior." We agree that "being emotional" often leads to bad judgments and undesired consequences, but still consider emotion a necessary component of advanced intelligence. Of course, the emotions in NARS are not "human-like" in details, but play similar roles as in human cognition, that is, situation appraisal and behavior control.

Due to AIKR, NARS is not aware of all of its internal structures and processes, but only the most prominent parts, "the tip of an iceberg." Most activities within the system are beyond the scope of self-awareness, so cannot be deliberately considered. The picture is like what Freud (1965) drew about human thinking, although in NARS the unconscious processes follow the same logic as the conscious processes, except unnoticed by the system's limited attention.

In general, NARS treats its "external experience" and "internal experience" in the same way, and the knowledge about the system itself has the same nature as other knowledge in NARS. Under AIKR, self-knowledge is incomplete, uncertain, and often inconsistent, which is the contrary of what is assumed by the "logical AI" school (McCarthy, 1995). The system can only be aware of the knowledge reported by certain mental operations and those in the input buffers, and even this knowledge does not necessarily get enough attention to reveal its implications.

4.3. Self-Control

Although the system only has limited self-knowledge, it nevertheless make self-control possible.

First, it is necessary to clarify what "self-control" means in this context. As almost all control activities are carried out by the system and the results are often within the system, to consider all of them "self-control" would trivialize the notion. Instead, the label should be limited to the actions resulting from the system's case-by-case reflective and introspective deliberation, rather than from the working routines that are in the system's initial design, as the latter should not be considered as the decision "by the system itself," but by the designer of the system.

A widely agreed conclusion in psychology is that a mental process can be either *automatic* (implicit, unconscious) or *controlled* (explicit, conscious), with respect to the system itself. The former includes innate or acquired stimulus–response associations, while the latter includes processes under *cognitive control*, such as "response inhibition, attentional bias, performance monitoring, conflict monitoring, response priming, task setting, task switching, and the setting of subsystem parameters, as well as working memory control functions such as monitoring, maintenance, updating, and gating" (Cooper, 2010). Various "dual-process" models have been proposed in psychology to cover both mechanisms, such as the study by Kahneman (2011). Such models are also needed in AI, even though the purpose here is not to simulate the human mind in all details, but to benefits from the advantages of both. In general, controlled processes are more flexible and adaptive, while automatic processes are more efficient and reliable. Such a model often uses meta-level processes to regulate object-level processes (Cox, 2005; Marshall, 2006; Shapiro and Bona, 2010; Rosenbloom et al., 2016), and such works are also covered in the study of machine consciousness (Chella et al., 2008; Baars and Franklin, 2009).

Even though this "object-level vs. meta-level" distinction exists in many systems, the exact form of the boundary between the two levels differs greatly, partly due to the architectures involved. A process should not be considered "meta" merely because it gets information from another process and also influences the latter, since the relation can be symmetric between the two, while normally the object-level processes have no access to the meta-level processes.

As a reasoning system, in NARS, "control" means to select the premise(s) and the rule(s) for each inference step, so as to link the individual inference steps into task-processing processes. The primary control mechanism of NARS is coded in a programming language and is independent of the system's experience. It is automatic and unconscious, in the sense that the system does not "think" about what to do in each step, but is context driven and data driven, while the data involved come from selections biased by dynamic priority distributions. On top of this, there are mental operations that are expressed in Narsese and invoked by the system's decisions, as a result of "conscious" inference activities. This meta-level deliberative control does not change the underlying automatic routines, but supplement and adjust them. This design is different from the metacognition implemented in the other systems (Cox, 2005) in that the operations in NARS are light weight and can be accomplished within a constant time, rather than decision-making procedures that compare the possible actions in detail with a high computational cost. In this aspect, they are similar to the "mental acts" in GLAIR (Shapiro and Bona, 2010).

Like the situation of self-awareness, in NARS, self-control is far from "complete" in any sense, because of AIKR. The system can only make limited adjustments in its control mechanism, so cannot "completely reprogram itself" and nor can it guarantee the absolute correctness of its self-control decisions, as they are based on the experience of the system, while the future can be different.

4.4. Self-Organization

There are processes in NARS where the *SELF*-concept and mental operations are not directly involved although the related issues are usually involved in the discussions related to "self."

One natural expectation for AI systems is that their functions and capabilities should not be completely "handcrafted," but self-constructive and self-organizing (Simon, 1962; Thórisson, 2012). We share this opinion, and therefore in NARS, "self-organization" and "learning from experience" refer to the same group of activities, which happens in various aspects of the system:

- **Knowledge.** According to experience-grounded semantics, the "knowledge" of NARS is not an objective description of the environment, but a summary of the system's subjective experience. The sensory experience is restricted by the system's sensors and its social experience by its linguistic capability and communicational channels. Furthermore, the system does not merely remember whatever it has experienced, but selectively keeps them, and generates conclusions and concepts to summarize and generalize the experience, so as to deal with new situations efficiently. NARS is not a traditional "symbolic system" that merely refers to the objects and events existing outside. Instead, the concepts and statements capture the regularities and invariants in its experience, so are fundamentally from the view point of the system itself. For an object, what the system knows is not its objective characters, but is "affordance" to the system, using the vocabulary of Gibson (1986).
- **Skill.** A special type of knowledge is the *skills*, i.e., procedural knowledge guiding the usage of the system's operations. As described previously, each operation is evoked when a certain condition is satisfied, and compound operations can be formed. Although some of such knowledge is innate, similar to the primitive reflexes of human beings, they nevertheless can be modified by the system's experience. Among all possible compounds, which ones will be actually formed also depends on the system's experience, like skill acquisition in humans. NARS has the ability of self-programming, in the sense that the system can organize its atomic operations into compound operations recursively and use them as a whole, so as to avoid repeated planning or searching (Wang, 2012b). In this aspect, NARS is similar to the "recursive self-improvement" model in the study by Steunebrink et al. (2016).
- **Motivation.** The motivational structure of the system is under constant adjustments and developments and is not fully specified by its designer or users. NARS is built to accept any task expressible in Narsese in any time, although the priority of each task will be adjusted by the system, and the system may even ignore some given tasks, as the consequences of conflict resolution, preemptive action, redundancy reduction, etc. From the given tasks and the system's beliefs, derived tasks are generated recursively *via* backward inference, initially as means to achieve the given tasks, but may gradually become autonomous. As the system "grows up," its motivational structure gradually evolves, and all the tasks in it collectively decide what the system desires at the moment. Therefore, the goals and drives of the system are determined by the system's design, the given tasks, and the experience of the system, but not by any of these factors alone (Wang, 2012a).

In summary, there is a relatively clear distinction between *object-level* and *meta-level* in NARS, where the former is specified in Narsese and formed *via* self-organization, while the latter is specified in the programming language (such as Java) and mostly independent of the system's experience.

Since all aspects of the object-level can be learned, everything expressible in Narsese is learnable, in the sense that it can be entered into the system, derived by the inference rules, as well as modified by new experience. Consequently, NARS is more sensitive to its experience than most AI systems developed so far, and learning happens in several different forms in various parts of the system. This treatment of learning is fundamentally different from the current machine learning paradigm (Russell and Norvig, 2010; Flach, 2012), since in NARS the learning processes do not follow algorithms and nor do they necessarily produce problem-specific mappings (Wang and Li, 2016).

This sensitivity to experience does not mean pure subjective or arbitrary behaviors. The objectivity in knowledge comes from communication and socialization. Generally speaking, the more a NARS-based system communicates with other systems and humans, the more objective it usually becomes, and the less its idiosyncratic experience matters, because its beliefs are based more on the common experience shared by the community it belongs to, although it is hard, if not impossible, to quantify this "extent of objectiveness."

On the other hand, in NARS, the meta-level knowledge is built into the system and immune to experience-triggered modification. This level includes the grammar rules of Narsese, the inference rules of NAL, the basic routines of memory management and inference control, the set of mental operators, etc. Even taken self-awareness and self-control into consideration, this built-in core is still fixed. As stated in the study by Hofstadter (1979), "Below every tangled hierarchy lies an inviolate level." Some approaches of recursive self-improvement suggest more radical and thorough self-modifications, but they usually ignore AIKR by assuming that the system can be sure that its self-modification can really improve its performance and that the system can afford the computational cost of complex deliberation and modifications needed for such improvements (Schmidhuber, 2007; Goertzel, 2014). We consider such assumptions unrealistic and therefore is irrelevant to the design and development of AGI systems.

4.5. Consciousness

Among the issues related to "self," *consciousness* is probably the most confusing one. This topic can be addressed from many different perspectives (Blackmore, 2004), and there is still less consensus on its basic form and function. Many people consider it impossible in AI, although there have been attempts to produce consciousness in computers (Baars and Franklin, 2009) or robots (Chella et al., 2008), based on various interpretations of the notion.

Here, we focus on the so-called hard problem, that is, how physical processes in the brain give rise to subjective experience (Chalmers, 1996). Our position, briefly speaking, is that the problem is not between "physical process" and "subjective experience" but between different types of experience.

As explained previously, the experience-grounded semantics (EGS) of NARS defines truth value of statements and meaning of concepts according to the system's experience and therefore rejects the assumption of an "objective description" of the world that is independent of any observer. Although the *world* (or call it "environment," "universe," etc.) exists independently of any observer, a *description* of it does not. First, a sensation is produced by a sensor; then, a perception depends on the generalization and association capability and the available concepts of the observer; finally, when the perception eventually becomes a description, the system must have paid enough attention to it, which in turn demands a relevant motivation, a proper emotional status, and so on. Therefore, there is no description that is from the viewpoint of nobody and describes the world "as it is." The so-called objective description is nothing but the shared opinions among human beings formed from communication, socialization, education, and so on, so it is not from any single person's viewpoint, but that of a human society. Therefore, this "objective" is actually "intersubjective" (Gillespie and Cornish, 2009). Beside the culture heritage, our descriptions of the world heavily depend on the common sensorimotor mechanism of the human species, which is not necessarily shared by all cognitive systems, like the other animals or robots, either the existing ones or the future ones.

Nagel (1974) raised the question of "What is it like to be a bat?," which has an obvious analogy in AI, "What is it like to be a robot?" As the sensorimotor mechanisms of robot are not identical to those of the human beings, we should not expect them to form concepts whose contents are exactly the same as human concepts, although through communication with human, shared concepts with overlapping meaning are possible to various extents, depending on the design of the robot and its training and working environment. This conclusion is not limited to robots. Actually EGS can be applied to any system, as far as it has interaction without its environment. For such a system to become "grounded," "embodied," or "situated," the key is not whether its input/output mechanisms are "human-like," but whether its behaviors depend on its experience (Wang, 2009).

A direct implication of the above conclusion is that intelligent systems in the same world may form different descriptions of the world, due to their different sensorimotor organs, concept repositories, motivational orientations, etc., even when their cognitive mechanisms are basically the same. In this situation, all these descriptions are valid, even when they are incommensurable. This is not saying that any arbitrary description is valid, but that its validity can only be evaluated according to the system's configuration and experience, rather than according to "the facts."

The same is true within the same system. If the system applies two different sets of sensorimotor mechanisms to the same process, it may get two descriptions, which are correlated, but incommensurable, and cannot be reduced into each other. We believe that this is exactly where the "explanatory gap" comes in consciousness.

As described above, NARS has internal experience about what is going on inside the system, which directly comes from the mental operations and the related introspective functions. When the system also learns how its own design works from a third-person perspective, even when it is given a way to observe its own running process at the machine language level, it will also have two incommensurable descriptions with a gap in between. In this case, it is incorrect to consider the high-level (mental) descriptions as "raised from" the low-level (physical) descriptions, as the latter is not "more real" than the former in some sense. This position also rejects the possibility of "zombies" that behave just like us, but have no consciousness (Chalmers, 1996), because if the system does not have internal experience, it will lack certain cognitive functions and therefore will not behave just like conscious beings.

In summary, we believe that the design of NARS enables the system to have consciousness, and the related phenomena can be explained without being reduced into phenomena in neuroscience (Koch, 2004) or quantum physics (Penrose, 1994). In AGI systems, although initially the conscious functions will be relatively simple and poor, they will become more and more complicated and rich, as the research progresses. The fact that we cannot directly sense them cannot be used to deny their experience, just like one cannot deny the consciousness of another person simply because one cannot directly know what it is like to be that person.

5. CONCLUSION

Self-awareness and self-control are important cognitive functions needed by advanced AGI systems (Chella and Manzotti, 2012). For a system to solve various types of problems, especially novel ones, it needs to know about itself, as well as to adjust its own working processes, so as to efficiently produce the best answer it can find with the current evidence and resource supply.

Just as a system's knowledge and control of its external environment are usually incomplete and fallible, so are its knowledge and control of its internal environment. An AGI system can learn how itself works using its introspective capability, especially the mental operations. It can also deliberately invoke some mental operations to realize the system's decisions and to adjust its working procedures. These functions enable the system to better adapt to its environment and to carry out its various tasks more efficiently. Even so, it can never fully know itself nor can it have complete self-control.

Although the study of self-awareness and self-control in NARS is still at an early stage, the conceptual design described above has been implemented, and is under testing and tuning. There are many details to be refined, and many self-related issues to be further explored, like those discussed in the studies by Hofstadter (1979) and Blackmore (2004). We believe the overall design is in agreement with the scientific knowledge on these processes in the human mind and also meets the needs and restrictions in AGI systems. We also believe that almost all self-related functions observed in the human mind will be reproduced in AGI systems in principle although the details will be different. Furthermore, these functions should not be modeled one by one in isolation, but all together according to the same basic principles of intelligence.

AUTHOR CONTRIBUTIONS

PW proposed the overall structure and drafted Sections 1, 4, and 5. XL drafted Section 2. PH drafted Section 3. All authors revised the whole article.

ACKNOWLEDGMENTS

Thanks to the anonymous reviewers for their helpful comments and suggestions. Thanks to the other members of our research team for proofreading the manuscript.

REFERENCES

Arbib, M., and Fellous, J.-M. (2004). Emotions: from brain to robot. *Trends Cogn. Sci.* 8, 554–559. doi:10.1016/j.tics.2004.10.004

Baars, B. J., and Franklin, S. (2009). Consciousness is computational: the LIDA model of global workspace theory. *Int. J. Mach. Conscious.* 1, 23–32. doi:10.1142/S1793843009000050

Bach, J. (2009). *Principles of Synthetic Intelligence PSI: An Architecture of Motivated Cognition*. Oxford: Oxford University Press.

Bieger, J., and Thórisson, K. R. (2016). “Artificial pedagogy: a proposal,” in *The Joint Multi-Conference on Human-Level Artificial Intelligence*. New York City: Doctoral Consortium.

Blackmore, S. (2004). *Consciousness: An Introduction*. Oxford: Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chella, A., Frixione, M., and Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artif. Intell. Med.* 44, 147–154. doi:10.1016/j.artmed.2008.07.003

Chella, A., and Manzotti, R. (2012). “AGI and machine consciousness,” in *Theoretical Foundations of Artificial General Intelligence*, eds P. Wang and B. Goertzel (Paris: Atlantis Press), 263–282.

Cooper, R. P. (2010). Cognitive control: componential or emergent? *Top. Cogn. Sci.* 2, 598–613. doi:10.1111/j.1756-8765.2010.01110.x

Cox, M. T. (2005). Metacognition in computation: a selected research review. *Artif. Intell.* 169, 104–141. doi:10.1016/j.artint.2005.10.009

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. New York, NY, USA: Cambridge University Press.

Franklin, S. (2007). “A foundational architecture for artificial general intelligence,” in *Advance of Artificial General Intelligence*, eds B. Goertzel and P. Wang (Amsterdam: IOS Press), 36–54.

Freud, S. (1965). *The Interpretation of Dreams*. New York: Avon Books. Translated by James Strachey from the 1900 edition.

Gibson, J. J. (ed.). (1986). “The theory of affordances,” in *The Ecological Approach To Visual Perception*, Chap. 8 (Hillsdale, New Jersey: Psychology Press), 127–143. new edition.

Gillespie, A., and Cornish, F. (2009). Intersubjectivity: towards a dialogical analysis. *J. Theory Soc. Behav.* 40, 19–46. doi:10.1111/j.1468-5914.2009.00419.x

Goertzel, B. (2014). GOLEM: towards an AGI meta-architecture enabling both goal preservation and radical self-improvement. *J. Exp. Theor. Artif. Intell.* 26, 391–403. doi:10.1080/0952813X.2014.895107

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach*. Greenwood Village: Roberts and Company.

Kowalski, R. (1979). *Logic for Problem Solving*. Amsterdam, The Netherlands: North-Holland publishing Co.

Luger, G. F. (2008). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 6 Edn. Boston: Pearson.

Marshall, J. B. (2006). A self-watching model of analogy-making and perception. *J. Exp. Theor. Artif. Intell.* 18, 267–307. doi:10.1080/09528130600758626

McCarthy, J. (1995). “Making robots conscious of their mental states,” in *Proceedings of Machine Intelligence 15, Intelligent Agents*, Vol. 15 (Oxford: Oxford University), 3–17.

Minsky, M. (1985). *The Society of Mind*. New York: Simon and Schuster.

Nagel, T. (1974). What is it like to be a bat? *Philos. Rev.* 83, 435–450. doi:10.2307/2183914

Peirce, C. S. (1931). *Collected Papers of Charles Sanders Peirce*, Vol. 2. Cambridge, MA: Harvard University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Piaget, J. (1963). *The Origins of Intelligence in Children*. New York: W.W. Norton & Company, Inc. Translated by M. Cook.

Poole, D. L., and Mackworth, A. K. (2017). *Artificial Intelligence: Foundations of Computational Agents*, 2 Edn. Cambridge: Cambridge University Press.

Rosenbloom, P. S., Demski, A., and Ustun, V. (2016). The Sigma cognitive architecture and system: towards functionally elegant grand unification. *J. Artif. Gen. Intell.* 7, 1–103. doi:10.1515/jagi-2016-0001

Russell, S., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*, 3rd Edn. Upper Saddle River, NJ: Prentice Hall.

Schmidhuber, J. (2007). “Gödel machines: fully self-referential optimal universal self-improvers,” in *Artificial General Intelligence*, eds B. Goertzel and C. Pennachin (Berlin: Springer), 199–226.

Shapiro, S. C., and Bona, J. P. (2010). The GLAIR cognitive architecture. *Int. J. Mach. Conscious.* 2, 307–332. doi:10.1142/S1793843010000515

Simon, H. A. (1962). “Artificial intelligence and self-organizing systems: experiments with a heuristic compiler,” in *Proceedings of the 1962 ACM National Conference on Digest of Technical Papers*.

Steunebrink, B. R., Thórisson, K. R., and Schmidhuber, J. (2016). “Growing recursive self-improvers,” in *Proceedings of the Ninth Conference on Artificial General Intelligence*, New York, 129–139.

Thórisson, K. R. (2012). “A new constructivist AI: from manual methods to self-constructive systems,” in *Theoretical Foundations of Artificial General Intelligence*, eds P. Wang and B. Goertzel (Paris: Atlantis Press), 145–171.

Wang, P. (2001). “Confidence as higher-order uncertainty,” in *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications* (Ithaca, NY), 352–361.

Wang, P. (2005). Experience-grounded semantics: a theory for intelligent systems. *Cogn. Syst. Res.* 6, 282–302. doi:10.1016/j.cogsys.2004.08.003

Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.

Wang, P. (2008). “What do you mean by ‘AI’,” in *Proceedings of the First Conference on Artificial General Intelligence*, Memphis, 362–373.

Wang, P. (2009). "Embodiment: does a laptop have a body?" in *Proceedings of the Second Conference on Artificial General Intelligence*, Arlington, Virginia, 174–179.

Wang, P. (2012a). "Motivation management in AGI systems," in *Proceedings of the Fifth Conference on Artificial General Intelligence*, Oxford, United Kingdom, 352–361.

Wang, P. (2012b). Solving a problem with or without a program. *J. Artif. Gen. Intell.* 3, 43–73. doi:10.2478/v10229-011-0021-5

Wang, P. (2013). *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. Singapore: World Scientific.

Wang, P., and Hammer, P. (2015). "Issues in temporal and causal inference," in *Proceedings of the Eighth Conference on Artificial General Intelligence*, Berlin, 208–217.

Wang, P., and Li, X. (2016). "Different conceptions of learning: function approximation vs. self-organization," in *Proceedings of the Ninth Conference on Artificial General Intelligence*, New York, 140–149.

Wang, P., Talanov, M., and Hammer, P. (2016). "The emotional mechanisms in NARS," in *Proceedings of the Ninth Conference on Artificial General Intelligence*, New York, 150–159.

Conflict of Interest Statement: 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.

The reviewer, HD, and handling editor declared their shared affiliation.

frontiers in Robotics and AI

PERSPECTIVE
published: 03 April 2018
doi: 10.3389/frobt.2018.00032

In Situ Representations and Access Consciousness in Neural Blackboard or Workspace Architectures

*Frank van der Velde**

Cognitive Psychology and Ergonomics, Institute for Digital Society, University of Twente, Enschede, Netherlands

Phenomenal theories of consciousness assert that consciousness is based on specific neural correlates in the brain, which can be separated from all cognitive functions we can perform. If so, the search for robot consciousness seems to be doomed. By contrast, theories of functional or access consciousness assert that consciousness can be studied only with forms of cognitive access, given by cognitive processes. Consequently, consciousness and cognitive access cannot be fully dissociated. Here, the global features of cognitive access of consciousness are discussed based on neural blackboard or (global) workspace architectures, combined with content addressable or "in situ" representations as found in the brain. These representations allow continuous cognitive access in the form of a process of covert or overt queries and answers that could underlie forms of access consciousness. A crucial aspect of this process is that it is controlled by the activity of the *in situ* representations themselves and the relations they can initiate, not by an external controller like a CPU that runs a particular program. Although the resulting process of access consciousness is indeed based on specific features of the brain, there are no principled reasons to assume that this process cannot be achieved in robots either.

Keywords: access consciousness, connection paths, global workspace, *in situ* representations, neural blackboard architectures, robots

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Adam Linson,
University of Stirling, United Kingdom
Christian Balkenius,
Lund University, Sweden

****Correspondence:***
Frank van der Velde
f.vandervelde@utwente.nl

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *29 October 2017*
Accepted: *14 March 2018*
Published: *03 April 2018*

Citation:
van der Velde F (2018) In Situ Representations and Access Consciousness in Neural Blackboard or Workspace Architectures. Front. Robot. AI 5:32. doi: 10.3389/frobt.2018.00032

1. INTRODUCTION

In one sense, discussing consciousness in a (humanoid) robot is easier than discussing human consciousness. In the latter case, we are hampered by our own "first-person" perspective. We "know" what it means to be conscious because we experience it ourselves. However, this first-person perspective is a form of introspection, which is out of reach for scientific observation and discussion.

The influence of the first-person perspective is clear in the distinction between two different views on the nature of consciousness, known as phenomenal consciousness and functional (or access) consciousness (Block, 1995; Cohen and Dennett, 2011; Taylor, 2012). Phenomenal consciousness asserts that conscious experiences result from specific neural correlates in the brain. Examples of these, depending on the theory at hand, are recurrent connections in the brain (e.g., Block, 2007), specific "microactivations" distributed over the brain (Zeki, 2003), or "winning" coalitions of neurons that result in a conscious experience of the representation they instantiate (Crick and Koch, 2003). The key notion of phenomenal consciousness is that the neural correlates responsible for consciousness can be separated (dissociated) from all the cognitive functions we can perform, such as attention, language, and the like. That is, consciousness "overflows access" (Cohen and Dennett, 2011).

For robotics, this would mean that the search for robot consciousness is doomed. Unless we endow robots with the required neural correlates (as in hybrid forms of neuro-robots), robots cannot possess forms of consciousness.

However, as convincingly argued by Cohen and Dennett (2011), phenomenal consciousness is a view on consciousness that is outside the reach of science precisely because it assumes neural correlates of consciousness separate from neural correlates of cognitive functions. A consequence of this view is that theories of consciousness cannot be empirically verified or falsified (which would always depend on some form of behavior produced by some kind of cognitive process).

By contrast, theories of functional or access consciousness assert that consciousness can be studied only with forms of cognitive access, given by cognitive processes. Consequently, consciousness and cognitive access cannot be fully dissociated. Instead, any form of consciousness would require a cognitive architecture that would allow forms of functional access.

An influential proposal for such an architecture is the Global Workspace theory, which asserts that consciousness arises when representations enter the Global Workspace of the brain and (temporarily) one of them dominates its activation (e.g., Baars, 2002; Baars and Franklin, 2003; Wiggins, 2012). The perspective I address here is that this theory could indeed provide the basis for a cognitive and computational architecture for functional consciousness, provided it is combined with a key observation on the nature of representation in the brain. This observation concerns the notion that representations in the brain are "*in situ*," which entails that they operate (at least in part) always as the same representation in each instantiation of the cognitive processes in which they participate.

In this way, *in situ* representations differ fundamentally from representations as used in von Neumann architectures, in which representations are inert, stored in arbitrary locations under the control of a CPU. By contrast, cognitive processes based on *in situ* representations are controlled by these representations, and not by an outside controller like a CPU that runs a particular program. This could result in a continuous process of "queries and answers" (van der Velde, 2013), which could form the basis for forms of access consciousness.

In the following sections, I will describe the notions of *in situ* representations, functional consciousness, and their relation in more detail.

2. *IN SITU* REPRESENTATIONS

A striking feature of representations in human cognition, as argued here, is their content-addressable nature. In this way, a representation can be (re)activated by directly activating it or a part of it. This is different from a representation in computers, which is accessed by means of its address label (which is also true for files in Github, where address labels are derived from the content of the file). In this case, a list of address labels needs to be run through first to find the label.

The notion of content-addressable representation is at the basis of many theories of human semantic representation (but see below for a counter example) and was one of the main motivations for the rise of connectionism in the 1980s (Bechtel and Abrahamsen, 1991). For example, Hebb (1949) used content addressability as the basis for his notion of the "cell assembly" hypothesis of (concept) representations in the brain. According to this hypothesis, a cell (or neural) assembly of a concept develops over time by interconnecting those neurons in the brain that are involved in processing information and generating actions related to that concept. These assemblies could be distributed over (very) different parts of the cortex (and other brain structures), depending on their nature.

A more recent version of a similar model of content-addressable representation in the brain is the "hub and spoke" theory of semantic representation in the brain (Lambon Ralph et al., 2017). In this theory, based on behavioral and imaging studies, modality-specific semantic information is represented in brain areas that process that kind of information (e.g., visual information in the visual cortex and auditory information in the auditory cortex). These kinds of representations are the "spokes" of semantic representations in the theory. However, the spokes are interconnected in (bi-lateral) hubs located in the anterior temporal lobes. Hub representations are transmodal, in that they respond to and correspond with cross-modal interactions of modality-specific information. Examples of transmodal representations in the temporal cortex were also observed in single-cell studies with human subjects. For example, neurons were found that responded to (the identity of) a person, regardless of whether the face of the person (visual information) or name (visual or auditory information) was presented (Quian Quiroga, 2012).

A crucial point here is that transmodal hub representations interconnect the modal spoke representations. But they do not replace or stand in for them. That is, their content is determined by the spoke representations they are connected to, and that content is reactivated when the hub representation is activated. This is what the cell assembly idea of Hebb is about. It is also in agreement with the imaging (fMRI) observations of Huth et al. (2016), who, in an extensive study, measured brain activity related to words when people were listening to stories. So, auditory language information was presented, but it activated a large set of cortical areas that responded to (also modal) semantic information, both in the left hemisphere (63 semantically selective areas) and the right hemisphere (77 semantically selective areas) of the cortex.

Hence, even though representations can have parts in transmodal hubs, they consist of a (potentially) large set of neurons (an assembly) distributed over widely different areas in the cortex, depending on their content. This shows why they are content addressable. By activating, say, the hub part of a representation, its spokes will be activated as well (as in the Huth et al. (2016)), revealing the content of the representation. But when (a part of) the spokes are activated by, for example, perceptual information, the hub part and consequently the other spokes can be activated as well. So, each activation of a representation potentially entails the activation of the entire hub and spokes. Crucially, this will be the same hub and set of spokes for each new activation of the representation (which, or course, can develop and change over time).

This is why these representations can be referred to as "in situ" (van der Velde, 2016). They do not consist of some (neural) code that can be copied and transported elsewhere, but of the entire web-like hub and spoke structure (which would be impossible to copy and transfer to somewhere else in the brain, e.g., given its distributed nature).

3. COMPUTATIONAL ARCHITECTURES BASED ON *IN SITU* REPRESENTATIONS

The nature of *in situ* representations in the brain raises the question of how they function in cognitive processes. In particular, in productive forms of cognitive processing, because these would seem to be the kind of cognitive processes that are needed to test forms of functional consciousness (Cohen and Dennett, 2011).

Productive processing entails that information is processed or produced in a combinatorial manner, based on (more elementary) constituent representations (concepts) and their relations. Productive processing is of key importance for human cognition, as found in language, reasoning, and visual perception. Consequently, they can be expected to play a key role in consciousness as well, as in relating conscious experiences to each other (e.g., *the apple is red* versus *the apple is green*).

Combinatorial processing with representations that are not copied but remain *in situ* can be achieved in architectures that provide (temporal) connection paths between the constituent representations, in line with their relations. For example, consider the combination *red apple*, with *in situ* representations for *red* and *apple*. Each one consists of an assembly structure with spokes in parts of the cortex related to perception or actions, such as seeing or eating an apple, and links to the transmodal hub in the anterior temporal cortex.

In the neural blackboard architecture of van der Velde and de Kamps (2006), the relation *red apple* is produced by establishing a (temporal) connection path between the *in situ* concept representations of *red* and *apple*, as illustrated in **Figure 1**. The path is achieved in a "neural blackboard," which could be connected in particular to the hub part of the *in situ* representations.

In this neural blackboard, the concepts are temporarily bound to "structure assemblies" in line with their word type. So, *apple* is bound to a "Noun assembly" in a "Noun field" and *red* is bound to an "Adjective assembly" in an "Adjective field." Such word type fields are in line with the existence of (agent and object) areas in the (temporal) cortex that are selectively activated when nouns function as agents (subjects) or objects of verbs (Frankland and Greene, 2015). In turn, the Noun assembly bound to *apple* and the Adjective assembly bound to *red* can be temporarily bound to each other, representing the relation *red apple*. The structure of the neural blackboard is such that it allows the combination of arbitrary words in a familiar language (van der Velde and de Kamps, 2015).

Neural blackboards would not only exist to process or produce conceptual structures (e.g., relations between words in a sentence) but also, for example, to process relations between visual features, as found in the structure of the visual cortex. Here, I am not discussing the specific way in which conceptual or visual relations can be processed in terms of *in situ* representations (see van der Velde and de Kamps (2006) for detailed descriptions), but the consequence of this form of representation for cognition, and potentially for functional consciousness, as outlined in the next section.

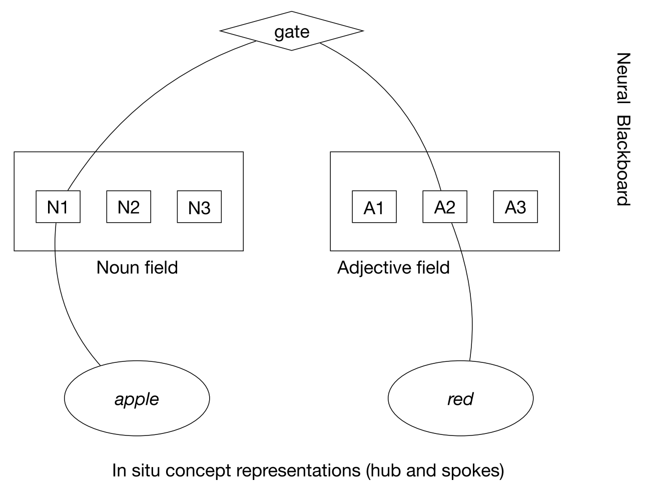


FIGURE 1 | *In situ* representation of *red apple* by a connection path between *in situ* concept representations in a neural blackboard architecture. The noun *apple* first binds to a Noun assembly (here, N1) in the Noun field of the neural blackboard and *red* binds to an Adjective assembly (here, A2) in the Adjective field. The connection path passes through gates, which provides control to represent relations. Here, activation of "gate" gives adjectives bound to nouns, here *red* to *apple*.

4. FUNCTIONAL CONSCIOUSNESS

The relation between *in situ* representations and functional consciousness, and they way they differ from phenomenal consciousness, can be illustrated with a "perfect" experiment described by Cohen and Dennett (2011). Assume we have a subject in which the area in the brain responsible for color consciousness is isolated from other brain areas higher up in the activation stream. So, this area (say V4 or inferotemporal cortex) would receive feedforward input from lower areas in the visual cortex, as in the normal situation, but cannot generate output to other areas. When a colored object is presented, say a red apple, the color area would be activated by and in correspondence with the color of the apple, as in the normal case. But activation of the color area itself is isolated from the rest of the brain.

Because of its isolation, the color area would not, for example, activate brain areas underlying language anymore. So, when presented with a red apple, our subject would not be able to say that the color is red. Indeed, she could not indicate by any form of action what the color of the apple is. Also, she would not become emotionally affected by the color (if that was the case before the isolation) because these areas are not activated by the color area anymore either. In fact, she would indicate that the apple is colorless. Yet, according to phenomenal consciousness theories, our subject would still be conscious of the color red, as its neural correlate is active. This activity could also be measured by brain imaging, supporting the notion that our subject is in fact conscious of red, even though she indicates by any form of action or emotion that she is not.

Thus, although our subject would not (and could not) indicate that she has a first-person experience of red, theories of phenomenal consciousness would still assert she has. But this assertion is untestable, because no action of our subject can indicate that she is conscious of it. Consequently, the theory that supports such a form of consciousness is unverifiable. To make a theory of consciousness verifiable, some form of action is needed to identify an experience as conscious. Hence, consciousness and cognitive functions are not fully dissociated. This is the notion underlying functional (or access) consciousness (Cohen and Dennett, 2011).

The "perfect" experiment of Cohen and Dennett (2011) relates directly to *in situ* representations because it entails that the *in situ* representation of, say, the concept *red* is broken. The *in situ* representation of this concept not just consists of the connections that activate it but also of the connections that activate related concepts and circuits that produce behavior related to the concept (as saying the word red or pointing to a red object in a display).

So, the integrity of an *in situ* representation, in particular its ability to produce behavior, is crucial for its role in functional consciousness. This raises a reverse question. Suppose a subject would produce behavior like saying that she is conscious of the color red. Is that sufficient to conclude she is? At face value, this would be the result of a theory of access consciousness, because the cognitive function of identifying the color would entail the conscious experience of it. In that case, all that would be required for robots to be conscious, say of colors, is their ability to indicate the color of an object, by speech or another cognitive function.

However, just saying "red" does not indicate what a subject is (fully) conscious of. Words are labels to indicate an experience or concept but often do not cover their entire content. This observation does not entail a form of phenomenal consciousness. It just indicates that more elaborate forms of access (other words, or other forms of action) are needed to unravel the content of a conscious experience.

To see how this could proceed, we need to look at the way in which *in situ* representations function in a Global Workspace architecture.

5. CONSCIOUSNESS BASED ON QUERIES AND ANSWERS WITH *IN SITU* REPRESENTATIONS

In the Global Workspace theory of consciousness, representations compete to get access to the workspace and to (temporarily) dominate it (e.g., Baars, 2002; Baars and Franklin, 2003; Wiggins, 2012). This raises the question of how representations can enter the workspace and how the domination of the workspace is related to consciousness. The notion that representations in the brain are *in situ* could provide the beginning of an answer to these questions. If so, the underlying architecture could also form a basis for robot consciousness.

An *in situ* representation would not "enter" the global workspace but instead would be connected to it, with a connection path as illustrated in **Figure 1**. If the workspace would have the structure of a neural blackboard as illustrated in this figure (or this neural blackboard would be a part of it), the "entrance" of a representation in the workspace would consist of a temporal activation of this connection path to and in the workspace. Several *in situ* representations could then compete, resulting in one representation (and its connection path) temporarily dominating the workspace.

The dominating *in situ* representation selected in the workspace could then form the basis for a functional form of consciousness by a (continuous) "process of explicit or implicit queries and answers" (van der Velde, 2013).

As an illustration, consider the entire representation of *red apple* in the neural blackboard architecture outlined in the previous section. Again, assume that a similar connection path would exist in the global workspace (or, alternatively, that the neural blackboard is a part of the workspace). Because of its *in situ* nature, the neural representation of the concept *red* would be connected to the visual areas in the brain that process and represent color, but also to the neural word representation *red* in language areas. The *in situ* representation of *apple* would be connected to the visual areas responsive to shape, and the word representation *apple* in the language areas.

The connection path between them in the neural blackboard (or global workspace) forms the basis for functional access and behavior, in which the relation between the *in situ* representations can be expressed in an action. So, for example, the (explicit or implicit) query "What is the color of the apple?" would be answered by activating the *in situ* representation of *apple* (e.g., by seeing it or hearing the word *apple* in an actual question) and the condition that allows the activation of Adjective assemblies bound to Noun assemblies (e.g., of *apple*) in the neural blackboard. In turn, this results in the activation of the *in situ* representation of *red* through the connection path that interconnects *apple* and *red* in the neural blackboard (or global workspace). This would form the basis for generating a response (reflecting functional access) such as pointing to the red object or reporting the word *red*.

The key notion of this process is that it is initiated and controlled by the *in situ* representations, and not by an outside controller like a CPU that runs a particular program. Hence, it will be a continuous process, in which activated *in situ* representations initiate queries to "ask" for other semantic information related to them (also represented by *in situ* representations). This continuous activation process underlies a continuous form of functional access, which in turn could be the basis for a process (stream) of access consciousness. More specific examples of this process and its relation to consciousness are presented in van der Velde (2013).

The importance of the fact that this process is controlled by *in situ* representations is further illustrated in **Figure 2**. This figure illustrates an indirect way of representing content information in the brain, as given by the indirection model of Kriete et al. (2013). Here, neural codes of *red* and *apple* are (temporarily) stored in "stripes" located in the prefrontal cortex (PFC). The stripes operate as registers in a computer memory. In turn, their address can be stored in other PFC "role stripes" (here for noun and adjective), which represents the relation *red apple*. So, the query *apple color?* can be answered by first retrieving the role stripe (here, noun-stripe) that stores the stripe address of *apple* (&Stripe1) and then going to the adjective stripe related to that noun-stripe. Then, the address code of the adjective (&Stripe2) can be retrieved, which will result in finding the location (stripe) where the neural code for the color (*red*) is stored.

In this process, the representations themselves are inactive and not content addressable. They can be retrieved only by finding the addresses of the locations where they are stored. These addresses can be different on different occasions, depending on the preceding representations stored in the process. As a result, the content of a given address (stripe) can vary from occasion to occasion. Hence, content and address are dissociated. So, activation of an address itself gives no information about its content and therefore cannot play a direct role in access consciousness.

Furthermore, a content code (e.g., of *red*) is generally not accessible when it is stored in a given stripe. For example, access to that given stripe needs to be blocked when other content is to be stored in other stripes. Otherwise, the content of the given stripe could inadvertently be deleted (overwritten) in the process of storing other representations in other stripes. Hence, the representation of *red* in **Figure 2** resembles the isolated color representation in the perfect experiment of Cohen and Dennett (2011) discussed earlier. It may be active within the stripe, but its access to processes outside the stripe, and hence its active involvement in these processes, is generally blocked. In other words, the content representations are generally inactive because the stripes in which they are stored are generally "closed."

The indirection model of Kriete et al. (2013) is a model for productive computing in the brain, closely resembling productive computing in a Von Neumann architecture. So the in-activeness (and in-accessibility) of representations and their (negative) consequences for functional or access consciousness as discussed earlier would also hold for the Von Neumann architecture, which underlies digital computing. In turn, digital computing still forms the basis of many robot systems, such as the iCub robot (Natale et al., 2016).

So, the analysis of access consciousness as given here would have consequences for robot consciousness as well. In particular, it would seem that forms of robot consciousness would require a computing architecture based on *in situ* computing as illustrated above, instead of the Von Neumann kind of architectures still used to date. If correct, robot consciousness would indeed be based on specific features of the brain. But, in contrast to the assertion of phenomenal consciousness, it would not be based on specific physiological features of the brain, most likely unobtainable for robots, but on its specific computing and cognitive architecture.

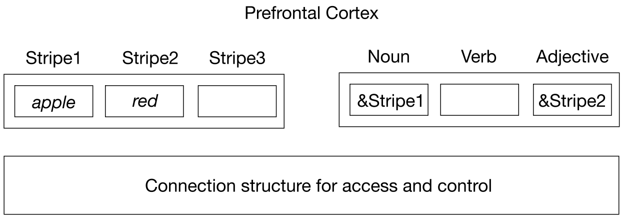


FIGURE 2 | Indirection representation of *red apple* based on Kriete et al. (2013). Neural codes for concepts (e.g., *apple*, *red*) are stored in memory locations ("stripes") in the prefrontal cortex. The addresses of these stripes (as given by the address operator &) are then stored in "role stripes," needed to establish the relation between the concepts. An underlying connection structure provides and controls access to stripes.

6. CONCLUSION ABOUT ROBOT CONSCIOUSNESS

The analysis presented here provides a few suggestions about the possibility and requirements of robot consciousness. First, consciousness seems to be related to *in situ* representations that underlie the possibility of cognitive access. Second, consciousness is more a process than a set of isolated conscious states. This, in combination with the requirement of access, suggests that consciousness is related to a continuous process of cognitive access. Third, this continuous process does not take the form of isolated instances of indirect activation of representations under the control of an external controller. Instead, the perspective is offered here that a continuous process of access can be achieved only when the process is directly controlled by (the activity of) *in situ* representations themselves, as in a continuous (covert or overt) process of queries and answers. Such a process seems to be in accordance with cognitive processing and access consciousness as found in the human brain. There are no principled reasons to assume that this process cannot be achieved in robots either.

AUTHOR CONTRIBUTIONS

FV conceived and wrote the article.

REFERENCES

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends Cogn. Sci.* 6, 47–52. doi:10.1016/S1364-6613(00)01819-2

Baars, B. J., and Franklin, S. (2003). How conscious experience and working memory interact. *Trends Cogn. Sci.* 7, 166–172. doi:10.1016/S1364-6613(03)00056-1

Bechtel, W., and Abrahamsen, A. (1991). *Connectionism and the Mind*. Cambridge, MA: Blackwell.

Block, N. (1995). On a confusion about the function of consciousness. *Behav. Brain Sci.* 18, 227–287. doi:10.1017/S0140525X00038188

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–499. doi:10.1017/S0140525X07002786

Cohen, M. A., and Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends Cogn. Sci.* 15, 358–364. doi:10.1016/j.tics.2011.06.008

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi:10.1038/nn0203-119

Frankland, S. M., and Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11732–11737. doi:10.1073/pnas.1421236112

Hebb, D. O. (1949). *The Organisation of Behaviour*. New York: Wiley.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi:10.1038/nature17637

Kriete, T., Noelle, D. C., Cohen, J. D., and O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16390–16395. doi:10.1073/pnas.1303547110

Lambon Ralph, M. A., Jefferies, E., Patterson, K., and Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18, 42–55. doi:10.1038/nrn.2016.150

Natale, L., Paikan, A., Randazzo, M., and Domenichelli, D. E. (2016). The icub software architecture: evolution and lessons learned. *Front. Rob. AI* 3:24. doi:10.3389/frobt.2016.00024

Quian Quiroga, R. (2012). Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* 13, 587–597. doi:10.1038/nrn3251

Taylor, J. G. (2012). Can functional and phenomenal consciousness be divided? *Int. J. Mach. Conscious.* 4, 457–469. doi:10.1142/S1793843012400264

van der Velde, F. (2013). Consciousness as a process of queries and answers in architectures based on in situ representations. *Int. J. Mach. Conscious.* 5, 27–45. doi:10.1142/S1793843013400039

van der Velde, F. (2016). Concepts and relations in neurally inspired in situ concept-based computing. *Front. Neurorobot.* 10:4. doi:10.3389/fnbot.2016.00004

van der Velde, F., and de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behav. Brain. Sci.* 29, 37–70. doi:10.1017/S0140525X06009022

van der Velde, F., and de Kamps, M. (2015). The necessity of connection structures in neural models of variable binding. *Cogn. Neurodyn.* 9, 359–370. doi:10.1007/s11571-015-9331-7

Wiggins, G. A. (2012). The minds chorus: creativity before consciousness. *Cognit. Comput.* 4, 306–319. doi:10.1007/s12559-012-9151-6

Zeki, S. (2003). The disunity of consciousness. *Trends Cognit. Sci.* 7, 214–218. doi:10.1016/S1364-6613(03)00081-0

Conflict of Interest Statement: 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.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 04 April 2018
doi: 10.3389/frobt.2018.00030

A Basic Architecture of an Autonomous Adaptive System With Conscious-Like Function for a Humanoid Robot

Yasuo Kinouchi and Kenneth James Mackin*

Department of Informatics, Tokyo University of Information Sciences, Chiba, Japan

In developing a humanoid robot, there are two major objectives. One is developing a physical robot having body, hands, and feet resembling those of human beings and being able to similarly control them. The other is to develop a control system that works similarly to our brain, to feel, think, act, and learn like ours. In this article, an architecture of a control system with a brain-oriented logical structure for the second objective is proposed. The proposed system autonomously adapts to the environment and implements a clearly defined "consciousness" function, through which both habitual behavior and goal-directed behavior are realized. Consciousness is regarded as a function for effective adaptation at the system-level, based on matching and organizing the individual results of the underlying parallel-processing units. This consciousness is assumed to correspond to how our mind is "aware" when making our moment to moment decisions in our daily life. The binding problem and the basic causes of delay in Libet's experiment are also explained by capturing awareness in this manner. The goal is set as an image in the system, and efficient actions toward achieving this goal are selected in the goal-directed behavior process. The system is designed as an artificial neural network and aims at achieving consistent and efficient system behavior, through the interaction of highly independent neural nodes. The proposed architecture is based on a two-level design. The first level, which we call the "basic-system," is an artificial neural network system that realizes consciousness, habitual behavior and explains the binding problem. The second level, which we call the "extended-system," is an artificial neural network system that realizes goal-directed behavior.

Keywords: goal-directed behavior, habitual behavior, autonomous adaptation, image processing, binding problem, Libet's experiment, model of consciousness, brain-oriented system

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Rodolphe Gelin,
Softbank Robotics, France
Salvatore Maria Anzalone,
Université Pierre et Marie Curie, France

****Correspondence:***
Yasuo Kinouchi
kinouchi@rsch.tuis.ac.jp

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *13 November 2017*
Accepted: *09 March 2018*
Published: *04 April 2018*

Citation:
Kinouchi Y and Mackin KJ (2018) A Basic Architecture of an Autonomous Adaptive System With Conscious-Like Function for a Humanoid Robot. Front. Robot. AI 5:30. doi: 10.3389/frobt.2018.00030

INTRODUCTION

Aims, Position, and Purpose of Research

In developing a humanoid robot, there are two major objectives. One is developing a physical robot having body, hands, and feet resembling those of human beings and being able to similarly control them (Jeffers and Grabowski, 2017; Tian et al., 2017). The other is to develop a control system that works similarly to our brain, to feel, think, act, and learn like ours (Dennett, 1994; Tani, 2017; Zorpette, 2017; Reggia et al., 2018). In this article, we propose an architecture as a basic logical structure of a brain-oriented control system toward realization of humanoid robot that feels, thinks, acts, and learns for the second objective. The reason for focusing on the architecture is that making the logical structure of the robot control system similar to our brain has the same important meaning as creating the physical structure of the robot resembling that of a human. The main behavioral characteristics of the humanoid robot will depend strongly on the basic logical structure of the control system.

To realize major operational characteristics of the brain in the system, we incorporate various findings from neuroscience and psychology to the proposed system. Knowledge on computer systems technology to realize highly complicated systems as well as latest artificial neural network designs is adopted at various levels to provide an integrated architecture.

Although the proposed robot's action is primitive by focusing on clearly defining the architecture, the control system of the robot has a function similar to consciousness and autonomously adapts to the environment. As an autonomous adaptation system, the robot feels, thinks, and learns through interactions with the environment. In addition, the duality of our behavioral characteristics—habitual behavior and goal-directed behavior—which has been the subject of research in a wide field including psychology and neuroscience (Deutsch and Strack, 2006; Kahneman, 2011; Mannella et al., 2016), is also realized in the control system by adopting a two-layer logical structure.

The model of consciousness included in the architecture clearly shows that consciousness is an essential function of the parallel-processing system and proposes the method of realizing consciousness and "self" from an engineering point of view. In addition, the model is positioned as improved model of global workspace theory (GWT) (Baars, 1988; Dehaene, 2014) and explains "unity," which is one of the basic characteristics of consciousness (Brook and Raymont, 2017).

The proposed architecture comprehensively accounts for the two major problems regarding consciousness still under debate, the time delay in Libet's experiment (Libet, 2004), and the binding problem (Feldman, 2013). This shows that the proposed architecture is not only valid as a brain-oriented architecture but also useful as a brain model from the viewpoint of information processing. Although the function level of the robot in this article is primitive, the proposed architecture can be applied to different problems and has high scalability. By expanding on the basic architecture, it will become possible to realize a humanoid robot with both mind and body. The architecture can be useful not only for humanoid robots but also for various types of autonomous robots, in general-purpose artificial intelligence (AI) development, and for understanding the brain.

Related Works, Methods, and Main Results

Recent developments in AI, particularly in deep learning, have shown remarkable achievements, such as mastering the game of Go (Silver et al., 2016), but current research is largely targeted toward particular fields and problems, and efforts toward brain-oriented design and human-like control systems are much smaller in comparison.

Even in the rapidly developing field of neuroscience, the whole brain's function as a control system has yet to be clarified. The neural mechanism behind "consciousness," a basic phenomenon of the brain, and "goal-directed behavior," the basis of everyday behavior, are still under debate (Gremel and Costa, 2013; Hart et al., 2013; Mannella et al., 2016). Human behavior is believed to be comprised of two distinct behavior characteristics, habitual behavior and goal-directed behavior, known as the duality of human behavior (Dezfouli and Balleine, 2013). Duality in human behavior has been widely studied in many fields, for example, fast and slow thinking by Kahneman (2011) in behavioral economics, and reflective-impulsive behavior model by Deutsch and Strack (2004, 2006) in psychology, but the basic neural mechanism has not been clarified.

We have previously proposed a conceptual control system that autonomously learns and makes behavior decisions based on primitive consciousness using an artificial neural network. We had proposed a model of consciousness as a system-level function and presented an artificial neural network system that enables fast decision of optimal behavior (Kinouchi and Kato, 2013; Kinouchi and Mackin, 2015). However, our previous proposal primarily explained only habitual behavior, and goal-directed behavior could not be explained yet.

On another front, various attempts have been proposed by Franklin et al. Haikonen has proposed the Haikonen cognitive architecture (HCA) and has been operating a robot with consciousness that adapts autonomously using a neural network (Haikonen, 2003, 2007, 2012). Franklin has been running a hybrid adaptation system, Learning Intelligent Distribution Agent (Franklin and Patterson, 2006; Franklin et al., 2013, 2014). In these, the method of action decision and the model of consciousness are both developed in accordance with Baars's proposed GWT (Baars, 1988; Baars and Franklin, 2007). In addition, Dehaene et al. proposed the global neuronal workspace that extended GWT from the viewpoint of neuroscience and tried to demonstrate it based on brain observation (Dehaene and Changeux, 2011; Dehaene, 2014). However, in these, perceptual filtering focused on only the most "salient" information is performed as an action selection based on GWT. As the salient information is not always optimal information for the system, the system's own profits is not strictly reflected in the action selection. We assume that the basis of action decision of autonomously adaptation system is to increase the profit of the system itself as much as possible at each time. Moreover, "self" that is an essential element of consciousness should correspond to the system itself trying to make the profit as large as possible.

First, we modified, reorganized and refined our previously proposed model as a core system for efficiently realizing habitual behavior (Kinouchi, 2009; Kinouchi and Kato, 2013; Kinouchi and Mackin, 2015). Hereafter, we call this core system the "basic-system." The basic-system autonomously adapts to the environment with functions of action decision based on profit optimization of the system at each time.

The main functions of the basic-system consist of primitive operations; (a) detecting objects from the environment and recognizing the objects, (b) action decision for the recognized objects, and (c) preparing next action including system-level learning. The importance of object handling function has been pointed out in the field of neuroscience, and then it is configured as a dedicated functional unit that enables the system to handle a bundle of signals, such as attributes of the object, collectively for processing. In action decision, an optimal action plan is calculated in a short time by using a recurrent neural network based on the Brain-State-in-a-Box (BSB), proposed by Anderson (1983) and Golden (1993). In addition, proposed circuit provides a function of powerful pattern match detection that detects matched pattern from thousands of parallel signals representing attributes of objects. This function is provided based on the findings related to pyramidal neuron (Spruston, 2008; Stuart and Spruston, 2015).

The basic-system is designed with priority on shortening response time and realized as a parallel-processing system that can quickly select desirable actions. To adapt itself to the environment, the system learns using an actor–critic reinforcement learning method, which is a kind of learning method without a teacher or a supervisor, under the control of evaluation unit incorporated in the system. Conscious phenomenon is regarded as activities for effective adaptation at the whole system-level, based on information integration and reconfiguration of individual results of the underlying paralleled functional units for preparing next action. The contents of consciousness are mainly composed of reconfigured information from attributes of the objects and evaluated value of the evaluation unit after action decision. And, these contents are transmitted to the related functional units in the whole system for speedy next action decision. These activities account for how our mind is "aware" when making our moment to moment decisions in our daily life.

Moreover, the binding problem and basic cause of the time delay in Libet's experiment is also explained comprehensively based on the above understandings for consciousness. In explaining both the binding problem and the Libet's delay, it is important that "the content of consciousness is reconfigured for the next action after action decision." Furthermore, for the binding problem, it is shown that functions handling bundled signals and a powerful pattern match detection functions also play an important role.

Next, to realize goal-directed behavior, we added functions for goal management to the basic-system. Hereafter, we call this enhanced system the "extended-system." In the extended-system, both habitual behavior and goal-directed behavior are comprehensively realized. The goal is represented and handled as a kind of object in the system, and efficient actions toward achieving the goal are successively executed.

In the extended-system, it is necessary to represent, to handle, and to recollect related reward and actions as well as the goal. To execute these functions effectively, the image handling functions are provided. In this article, we use the term "image" as "*information generated inside the system that the system can operate as an object (processing target)*" based on Haikonen (2003). Using these functions, it is possible to retrieve past experiences from long-term memory and refer to these contents for decision-making.

These operations are realized by repeated execution of the functions corresponding to the basic-system, aimed at higher reward acquisition over a long-time span. Here, consciousness is more than just "awareness" of a simple decision-making process but includes a kind of "will" or "intention" of the mind aiming at acquiring a higher level of reward, by processing sequential chains of multiple images.

BASIC CONDITIONS AND OUTLINE OF THE SYSTEM

Methods and Basic Conditions of the Control System

To grasp the fundamental logical structure of the brain as easily as possible, we adopt following method. First, we assume that "the brain is a kind of information processing system that satisfies certain conditions." Then, we clarify what functions are required, and what kind of logical configuration is necessary and efficient on the system satisfying the conditions. In this method, we do not directly imitate the structure of the brain or conscious phenomena. We expect that consciousness is designed or generated as one of the functions necessary for satisfying the system conditions; moreover, logical functions related to conscious phenomenon are totally included in the system. Based on the classification of Reggia (2013), our method is a kind of computational modeling of the "simulated consciousness" in a broad sense, but it also encompasses a part of the "instantiated consciousness." The validity or effectiveness of the logical structure is checked based on whether or not the major characteristics of the brain can be explained using the logical structure. As the main characteristics of the brain, consciousness, and related phenomenon, binding problem, delay of Libet's experiment, duality, etc., are used for validations.

Basic conditions of the control system are shown below.

(i) The control system autonomously adapts to the environment through learning. We consider that autonomous adaptation is the most fundamental and important system characteristic of the animal brain. To adapt itself to the environment without a teacher or a supervisor, the control system incorporates a functional unit that evaluates reward and punishment, acts under its own decision based on the evaluated value, and self-adapts based on the results of the action. As a humanoid robot control system, when the system receives a reward, the evaluation unit becomes a pleasant state, and on the other hand, when receiving a punishment, it becomes an unpleasant state. The degree of pleasant and unpleasant varies according to the degree of reward and punishment.

(ii) The system design is based on maximum performance and efficiency. The aim is not only to realize high performance but also to base the system design on maximum efficiency design. The assumption is that our brain is in a kind of optimal design state through natural selection process. By choosing maximum efficiency and optimal design from among various design possibilities, as a result, we expect that the selected design approaches that of the brain. Moreover, to realize many complexed functions with high performance, parallel processing is basically introduced.

(iii) The system is constructed by artificial neural networks. An artificial neural node is a processing element inspired by biological neural cells and is used as a basic computational element in deep learning and artificial neural networks. It is most effective from the viewpoint of high parallelism and flexible learning function. The processing speed of the element is assumed to be equivalent to an actual nerve cell.

Operating Environment of the Robot and Basic Configuration of the System

Because we prioritized understanding of the basic logical structure, the control system, the robot, and the environment are limited to indispensable functions or items and set as simple as possible. The robot and its operational environment are illustrated in **Figure 1A**. The robot has functions that detect objects, recognize the objects, approach or avoid the objects, and earn rewards or punishments through acquisition of the objects. The robot walks randomly when there is no object in sight. When one or more objects are captured, the robot selects one preferred object and acts for it. These behaviors are controlled by the control system in the robot head. (In the following, the control system is called "system.") Conceptual configuration of the system is shown in **Figure 1B**. The perception module detects and recognizes an object, and the action decision module determines an action, and the motor module executes the action. The memory module includes episodic memory. The system control module controls the operation of the whole system. Focusing on habitual behavior and goal-directed behavior, we designed the system in two stages as shown in **Figure 1C**. First, the basic-system realizes habitual behavior. Next, the extended-system, functional expansion of the basic-system, realizes goal-directed behavior.

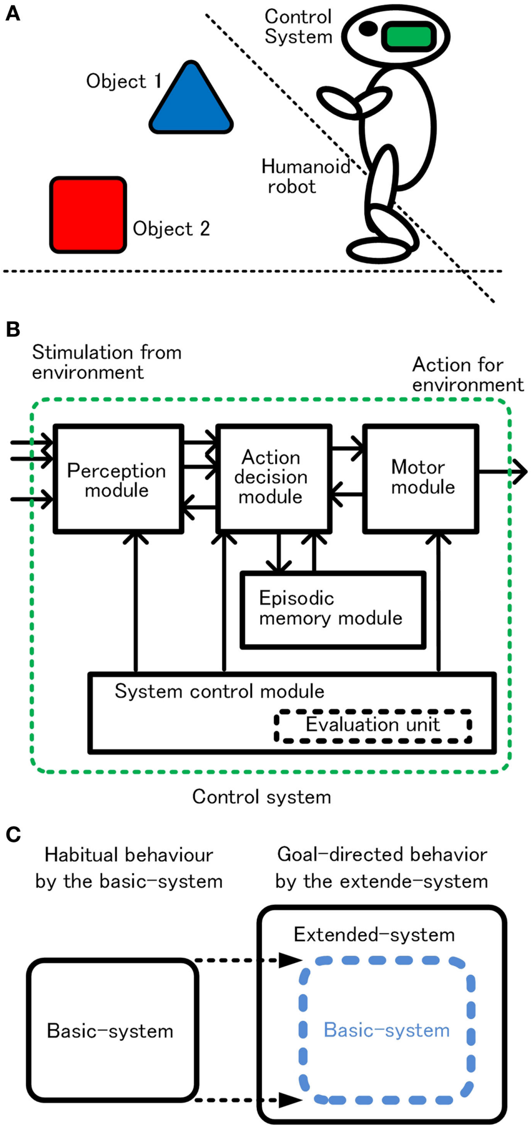


FIGURE 1 | Schematic configuration of the robot in the environment **(A)** and control system **(B)**, and two step approaches for goal-directed behavior **(C)**.

BASIC FUNCTIONS IN THE SYSTEM DESIGN

To configure an autonomous adaptive system using a neural network based on the basic conditions shown in the previous section, the following basic functions are further required.

a. Handling group of signals as a bundle and handling the bundle as an object.
b. Managing a signal as the signal with same meaning, even when used in various areas in the system.
c. Time management by the system itself and timing adjustment of a number of parallel operating functional units consistently.

In the case of computers, these functions are usually designed and implemented based on human designer. However, in the case of an autonomous adaptive system in which the system changes the system configuration itself, these functions must be implemented as basic functions in advance. On the premise of these functions, many dedicated functional units, such as recognition and action decision function, can operate in the autonomous adaptation system.

Handling a Group of Signals as a Bundle and Handling the Bundle as an Object

It has been reported that when animals or humans "perceive" something, inputs are selected from various stimuli to form an object, and then the object is later identified from detailed information and location information (Kahneman and Treisman, 1992; Pylyshyn, 2001; Xu and Chun, 2009). Object handling functionality has been reported to have a strong relationship with working memory features (Bays et al., 2011).

In computer systems, for effective operation, it is essential that the system can express and manage information composed of data that change over time, such as files and packets, as a bundle or a data set (Gray and Andreas Reuter, 1993; Patterson and Hennessy, 1994; Stalling, 2005). Various data or signals can be simultaneously exchanged or activated in a processor, but the data that can be processed by a program is limited to the data satisfying a specific condition, such as being on a general register or memory. For data satisfying the specific condition, a program can process the data regardless of whether that are data from an external source or internally generated data.

We have previously proposed the "object-handler" for bundling and handling information described earlier (Kinouchi and Mackin, 2015). In this article, we further clarify the functions of the object-handler for bundling signals, as well as using these bundles as an object. Only information maintained by the object-handler can be handled as an object regardless of where the signal originated from.

Management of Signal Meaning

Information that is widely used in the system must be interpreted with the same meaning throughout the system. In this article, for information that needs to be used over a wide area or over time, a node serving as a reference of the meaning of each signal is provided, thereby managing the meaning throughout the system. Hereafter, we call this node the "reference node." The meaning of each reference node is determined by the corresponding code conversion units, such as pattern and color recognition units.

When a functional unit uses signals whose meaning is already managed, the signal is supplied from the reference node. This method is based on the "*in situ*" representation proposed by van der Velde (2013). The system can be easily configured by allowing each unit to send and receive managed information bidirectionally from the reference node. In **Figure 2**, many functional units connected to the output of the reference node can receive signals at the same time. When one functional unit outputs a signal to the signal line, other functional units can receive the signal as a signal whose meaning is managed. In **Figure 2A**, a pair of unidirectional serial connection is provided for bidirectional transmission. In this way, signals whose meanings are managed by the reference node can be mutually transmitted and received between a large number of function units. Excitation of the reference node is unnecessary when merely transmitting and receiving the managed signals between the function units. Since this connection has a function similar to that of "bus" used in computers (Hwang and Briggs, 1984; Patterson and Hennessy, 1994), it is shown simplified as a bus in **Figure 2B**. When modifying the meaning of the reference node, a code conversion unit that determines the meaning of the reference node excites the reference node by the output of the code conversion unit.

System Time Management

In digital computers, the time adjustment between the functional units operating in parallel is controlled using a clock running at a constant rate with high accuracy (Patterson and Hennessy, 1994). However, it is difficult to adopt this method in the system. The reason is that the processing time of the functional unit in the system is not necessarily fixed and learning for adaptation may change the processing time of the unit itself. Nonetheless, for the system to operate parallel functional units satisfying the basic condition (ii) for high performance, timing adjustment between various units is essential.

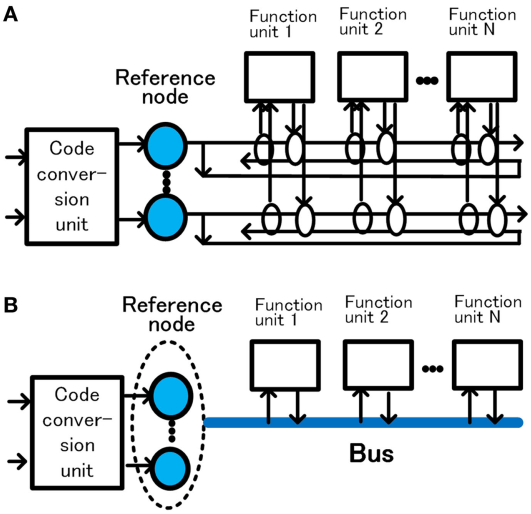


FIGURE 2 | Management of signal meaning by the reference nodes and bus. Configuration of a pair of unidirectional serial connection for bidirectional transmission **(A)** and configuration represented by bus **(B)**.

For basic timing adjustment, we used the case where the units in the system are excited simultaneously in wide-area mutual stimulation. The case indicates that the activation of each unit occurs at the same time and each unit can base the start timing from this signal. The excitement of the recurrent network for action decision described in Section "Decision Phase" provides this simultaneous excitation as a base point of timing and keeps the track of the system time by the number of iterations from this base point. However, the repetition time of this base point is long and not constant; the system subsidiarily uses together a constant period clock with short repetition time and low precision for a narrow time width. A method of dividing or slicing the clock time is also used to share the bus among various function units accessing the bus at the same time.

CONFIGURATION AND FUNCTIONS OF THE BASIC-SYSTEM

In this section, the configuration of the basic-system and how habitual behavior is realized by the basic-system based on the basic conditions is described. The configuration of the basic-system is shown in **Figure 3**. One processing cycle of the basic-system is composed of three phases, the preprocessing phase, the decision phase, and the postprocessing phase. Through repeated iteration of the processing cycle, habitual behaviors are executed as shown in **Figure 4**. In the preprocessing phase, the objects are detected, and in the decision phase, action for the object is decided. The instruction for action is issued immediately after the decision phase. In the postprocessing phase, the information in the system is reorganized and prepared for the next cycle. The reason for issuing an action instruction immediately after deciding an action is that fast response to a stimulus is a major feature of habitual behavior related to the basic condition (ii). For primitive animals, the length of response time to a stimulus often becomes a matter of life or death.

The basic time of the system is counted by the number of processing cycles. Since, many networks widely excite simultaneously for action decision, the basic point for time management of the system is set at the last point in each decision phase. The execution time of one processing cycle will be simply called a cycle hereafter.

Preprocessing Phase

Here, we describe the main process in the preprocessing phase of preparing information necessary for action decision, which consists of the following two steps:

a. Detecting information to be operated by the system and managing it as a bundle of information.
b. Executing pattern recognition and color recognition for bundled objects.

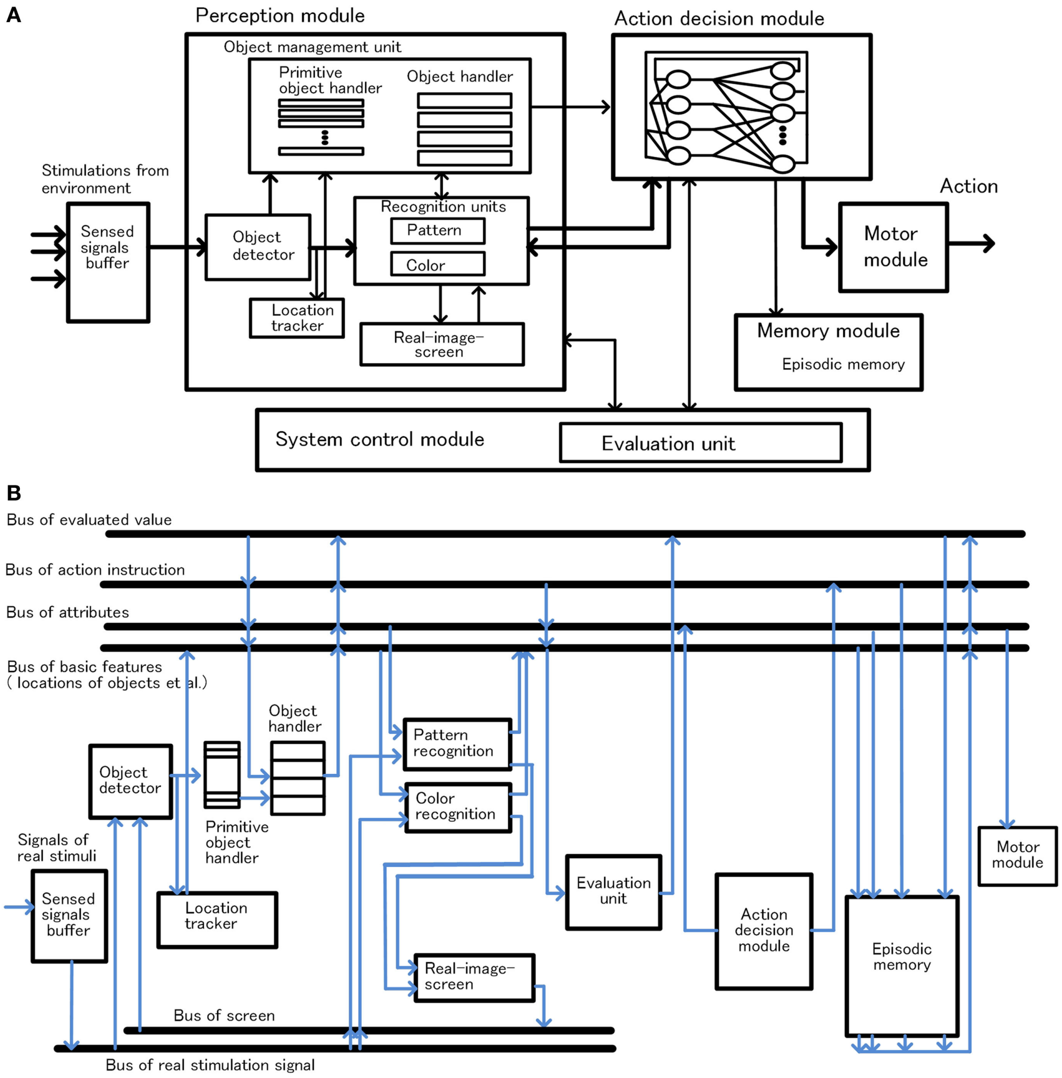


FIGURE 3 | Configuration of the basic-system, module configuration in panel **(A)** and bus configuration in panel **(B)**.

Object Detection and Management

Object detection and management are described according to **Figure 3**. When a group of stimuli generated in the sensed signals buffer, the object detector detects these signals as one bundle, and a primitive-object-handler in a free state captures it and sets it as a candidate of object. The primitive-object-handlers are functional units that maintain and manage temporary information of the candidate of object composed of sensed signals corresponding figure and location. From this point, the location of the candidate of object is tracked. Then, a free object-handler takes over the information of the candidate of object from the primitive-object-handler, and the object-handler starts management of the information as object. At the same time, the object-handler requests to recognize pattern or color of the object maintained in the object-handler to related functions. The object-handlers are functional units that maintain and manage the temporary information as bundles of information composed of sensed signals, location, and recognized attributes, such as pattern or color, of the object. [We assume that the primitive-object-handlers are related to function of the fragile memory, a kind of short-term memory, and the object-handlers are related to function of the working memory, based on Sligte et al. (2009, 2010); Scimeca et al. (2015); Block (2011) and Bays et al. (2011).]

Only the bundles of information managed by the object-handler can be processed for action decision by the system. This means that even if a bundle of information or signals is generated in the system itself, the bundle managed by the object-handler can be treated as an objective for action decision of the system. This method is applied to the image handling used in the extended-system. Details will be described later.

Object Recognition

Here, object recognition is described according to **Figure 5**. The object-handler instructs recognition units, such as pattern or color recognition, to recognize the object allocated to the object-handler, and maintains the attributes of the object as a result of

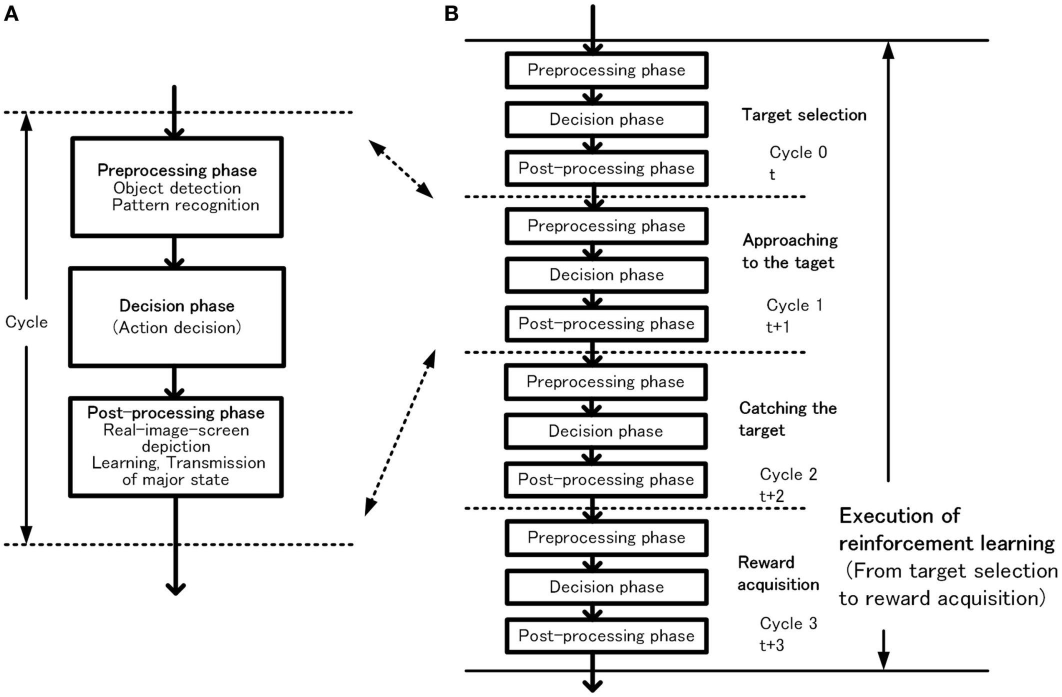


FIGURE 4 | Outline of habitual behavior in the basic-system, basic processing cycle **(A)** and cycles for execution of reinforcement learning **(B)**.

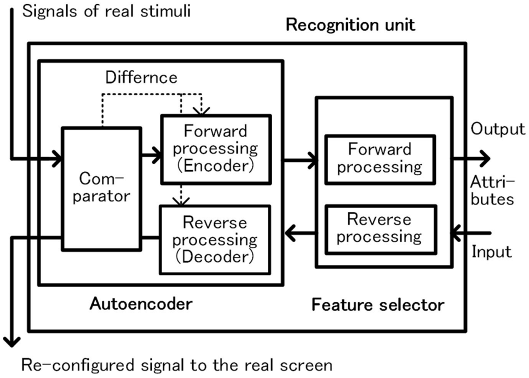


FIGURE 5 | Configuration of the recognition unit.

recognition mentioned earlier. In these operations, up to four object-handlers operate concurrently in the preprocessing phase considering the capacity of the working memory (Bays et al., 2011; Block, 2011).

The recognition unit is composed of a combination of autoencoder and feature selector as shown in **Figure 5**. The autoencoder extracts effective features for efficient expression of the sensed signals of the object, and then the feature selector specifies the features of each attribute. An important characteristic of this unit is that it operates bidirectionally. In forward processing, the unit recognizes the sensed signals as a pattern and outputs the attribute of the pattern. In backward processing, a group of attributes are input to the recognition unit from the opposite direction, and a pattern corresponding to the group of attributes is regenerated. In this case, the feature selector reproduces the feature group from the attribute pattern. Next, the autoencoder reproduces the input pattern based on the reproduced feature group. In the preprocessing phase, the recognition unit operates only in the forward direction, and in the postprocessing phase the recognition unit operates in the reverse direction. We have currently adopted a very simple recognition function. In the field of deep learning, which is rapidly developing in recent years, combination of autoencoder and feature selector is frequently used (Ranzato et al., 2007; Bengio, 2009; Larochelle et al., 2009). We expect that this method can be applied to improve the recognition function.

Decision Phase

Outline of the Decision Phase Operation

In the decision phase, satisfying the basic condition (ii), the system quickly selects the most desirable pair for the system at that time from a large number of objects and action pairs, and issues the result immediately as an action instruction using the recurrent neural network in **Figure 6**. The configuration of the recurrent neural network is equivalent to the BSB, proposed by Anderson (1983) and generalized by Golden (1986, 1993). Golden has revealed that the BSB is a gradient descent algorithm in the direction to reduce the cost represented by the cost function (corresponding to the energy function). BSB has been studied mainly as a method for categorization.

In this article, the cost function expressed by the quadratic expression of connection weights between nodes, corresponds to the desirability of the system (system desirability *D*). As shown in the following section, by changing the connection weights according to action evaluation, such as pleasant/unpleasant, the

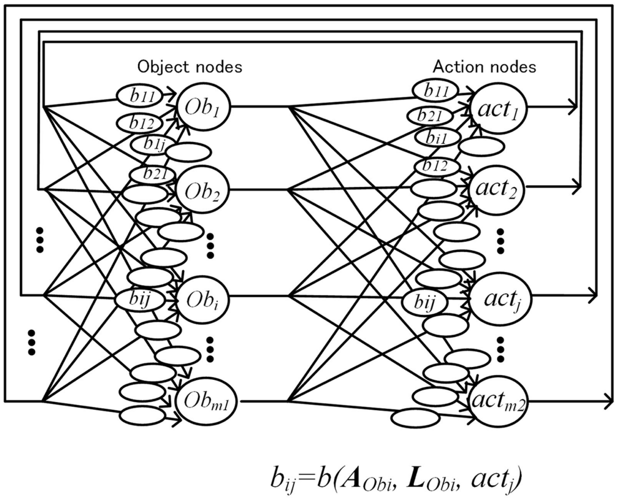


FIGURE 6 | Conceptual network configuration for action decision.

cost function can be modified and trained by the experience of the system. By performing the steepest descent algorithm under this cost function, optimization operation for the desirability of the system is possible using a recurrent neural network.

Detail Processing in Decision Phase

As shown previously, each object-handler maintains attributes and location information of the assigned object. Attribute of object i (Ob$_i$) is expressed by a vector $\boldsymbol{A}_{\mathrm{Ob}_i} = (a_1, a_2, \ldots, a_{k1})$. When Ob$_i$ has corresponding micro-feature j, then $a_j = 1$, and when Ob$_i$ has no corresponding micro-feature j, then $a_j = 0$. Similarly, location of Ob$_i$ is expressed by a vector $\boldsymbol{L}_{\mathrm{Ob}_i} = (l_1, l_2, \ldots, l_{k1})$. When Ob$_i$ is found at distance l_j, then $l_j = 1$, and when Ob$_i$ is not found at distance l_j, then $l_j = 0$. (For simplification, only one distance l_j is set to 1 and others are set to 0.)

The operation selecting the desirable object–action pair is speedily executed by iterations based on the BSB as shown in **Figure 6**. The cost function is defined by system desirability D as expressed in Eq. 1. Variables $x_{\mathrm{Ob}_i}(n,t)$ and $y_{\mathrm{act}_j}(n,t)$ represent the degree of how necessary or desirable object Ob$_i$ and action act$_j$ is for the system in the nth iteration at time t, and is implemented as the activation level of neural nodes, which correspond to Ob$_i$ or act$_j$. Coefficient $b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)$ indicates desirability of object–action pair of object Ob$_i$ and action act$_j$, and is implemented as the connection weights between object node i and action node j

$$D(n,t) = \sum_{\mathrm{Ob}_i,\mathrm{act}_j} b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)\, x_{\mathrm{Ob}_i}(n,t)\, y_{\mathrm{act}_j}(n,t). \tag{1}$$

Activation levels of object or action nodes are increased or decreased from initial states according to D in a limited number of iterations. After the iteration, detecting the object and action node with maximum activation means selecting the semi-optimum object–action pair for D at time t. In the optimization process, constraints such as $\sum_{\mathrm{Ob}_i} {x_{\mathrm{Ob}_i}}^2 \leq 1$ and $\sum_{\mathrm{act}_j} {y_{\mathrm{act}_j}}^2 \leq 1$ are applied, but for simplicity, these constraints are abbreviated in this article. Operations mentioned earlier are executed along the following equations. The characteristics of neural nodes are defined by Eqs 2 and 3 with a piecewise-linear activation function

$$x_{\mathrm{Ob}_i}(n+1,t) = f(\varphi_i(n,t)), \tag{2}$$

$$f(\varphi_i(n,t)) \begin{cases} = 1 & \text{if} \quad \varphi_i(n,t) > 1 \\ = \varphi_i(n,t) & \\ = 0 & \text{if} \quad \varphi_i(n,t) < 1 \end{cases}$$

where

$$\varphi_i(n,t) = x_{\mathrm{Ob}_i}(n,t) + \sum_{\mathrm{act}_j} b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)\, y_{\mathrm{act}_j}(n,t). \tag{3}$$

Equations 4–6 are lead from Eqs 1 to 3

$$\frac{\Delta D(n,t)}{\Delta x_{\mathrm{Ob}_i}(n,t)} \cong \sum_{\mathrm{act}_j} b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)\, y_{\mathrm{act}_j}(n,t)$$

$$x_{\mathrm{Ob}_i}(n+1,t) - x_{\mathrm{Ob}_i}(n,t) = \sum_{\mathrm{act}_j} b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)\, y_{\mathrm{act}_j}(n,t), \tag{4}$$

$$x_{\mathrm{Ob}_i}(n+1,t) - x_{\mathrm{Ob}_i}(n,t) \cong \frac{\Delta D(n,t)}{\Delta x_{\mathrm{Ob}_i}(n,t)}, \tag{5}$$

$$y_{\mathrm{act}_j}(n+1,t) - y_{\mathrm{act}_j}(n,t) \cong \frac{\Delta D(n,t)}{\Delta y_{\mathrm{act}_j}(n,t)}. \tag{6}$$

Based on above Eqs 4–6, the desirable object–action pair is selected using the gradient method in BSB.

The following two extensions are adopted for implementing the network to the basic-system:

A. The coefficient $b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)$ is effective only when an object of a certain attribute is in a certain place. This means that a single neural node must be able to detect patterns of attribute and location signals on its own. Previous artificial neural models require a large network of neurons for such pattern detection. To cope with this problem, we proposed a pattern match detection method inspired by the pyramidal neurons in the cerebral cortex, in which the dendritic structure support various matching detection. One pyramidal neuron has thousands of branches in the dendrite, and each branch processes thousands of paralleled input signals (Spruston, 2008; Kasai et al., 2010; Coward, 2013).

   Schematic diagram of the artificial neural node is shown in **Figure 7A**. Information is composed of main signal s_0 $(0 \leq s_0 \leq 1)$ and sub-signal $\boldsymbol{S}_a = (s_{a1}, s_{a2}, \ldots, s_{k3})$. For simplicity, $s_{ai} = 0$ or 1. Each branch memorizes a sub-signal pattern $\boldsymbol{S}_a$, where W_{S_a} is a weight corresponding to this pattern $\boldsymbol{S}_a$. This pyramidal neural node outputs $s_0 \cdot W_{S_a}$, only when input pattern $\boldsymbol{S}_a$ is matched with the pattern in the branches.

B. In the method shown in **Figure 6**, there is another disadvantage. As an object with a specific pattern of attributes and location is assigned to a fixed physical object node, the same object that has changed location is assigned as a different object. The object changed location should be treated as a

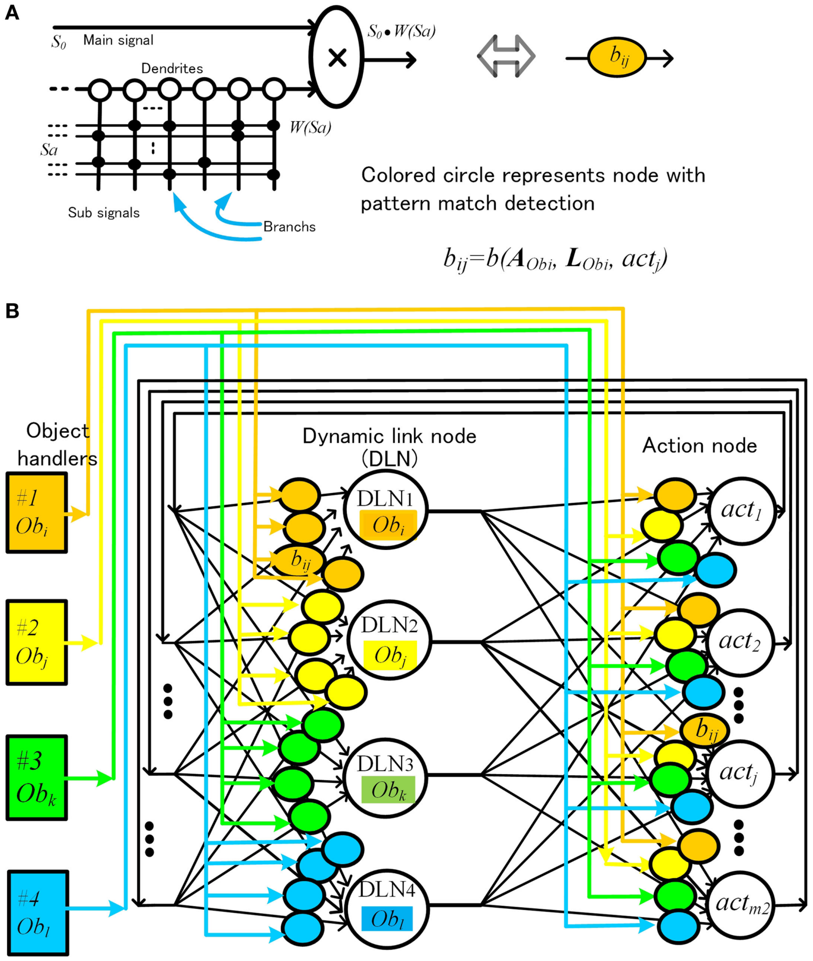


FIGURE 7 | Artificial neural node with pattern match detection **(A)** and schematic configuration of action decision network using dynamic link nodes **(B)**.

same object. To achieve this, we proposed a method called dynamic link node (DLN). Schematic configuration is shown **Figure 7B**. In this method, we limit the number of object nodes to 4, and we make four pairs of an object-handler and an object node. Each object node represents and functions as an object maintained by the paired object-handler. Each object-handler supplies attributes and location information to the paired object nodes. This means that same physical object node can operate as different object node dynamically by changing information maintained paired object-handler.

Equations 5 and 6 are transformed as below, corresponding to **Figure 7B**

$$r_k(\mathrm{Ob}_i, n+1, t) = r_k(\mathrm{Ob}_i, n, t) + \sum_{\mathrm{act}_j} b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)\, y_{\mathrm{act}_j}(n, t)\,,$$

$$y_{\mathrm{act}_j}(n+1, t) = y_{\mathrm{act}_j}(n, t) + \sum_{k(\mathrm{Ob}_i)} b_{ij}^t(\boldsymbol{A}_{\mathrm{Ob}_i}, \boldsymbol{L}_{\mathrm{Ob}_i}, \mathrm{act}_j)\, r_k(\mathrm{Ob}_i, n, t)\,.$$

where $r_k(\mathrm{Ob}_i, n, t)$ indicates activation of DLN k which work as node of Ob_i. $\boldsymbol{A}_{\mathrm{Ob}_i}$ and $\boldsymbol{L}_{\mathrm{Ob}_i}$ are supplied by the object-handler according to the object processed at that time. Although wired connection is fixed, the circuit in **Figure 7B** is able to process various objects dynamically.

However, implementing the circuits according **Figure 7B** is not easy. As the circuits have to wire four set of attributes and location signals to nodes, the circuit becomes very complicated. To avoid this problem, we introduced a time division method, controlled by a sub-clock, which sends four sets of attributes and position information using one set of wire. The configuration is depicted in **Figure 8**.

Postprocessing Phase

In the postprocessing phase, the system first reconfigures major information scattered in the system and performs necessary learning for adaptation of itself. Then, to respond quickly to new stimulus in the next cycle in line to the basic condition (ii), transmitting and processing of the major information are executed.

Reconfiguration of Information and Learning

The operation of the autonomous adaptation system can be described largely as two operations: (a) operations for external environment as an action of the system and (b) learning

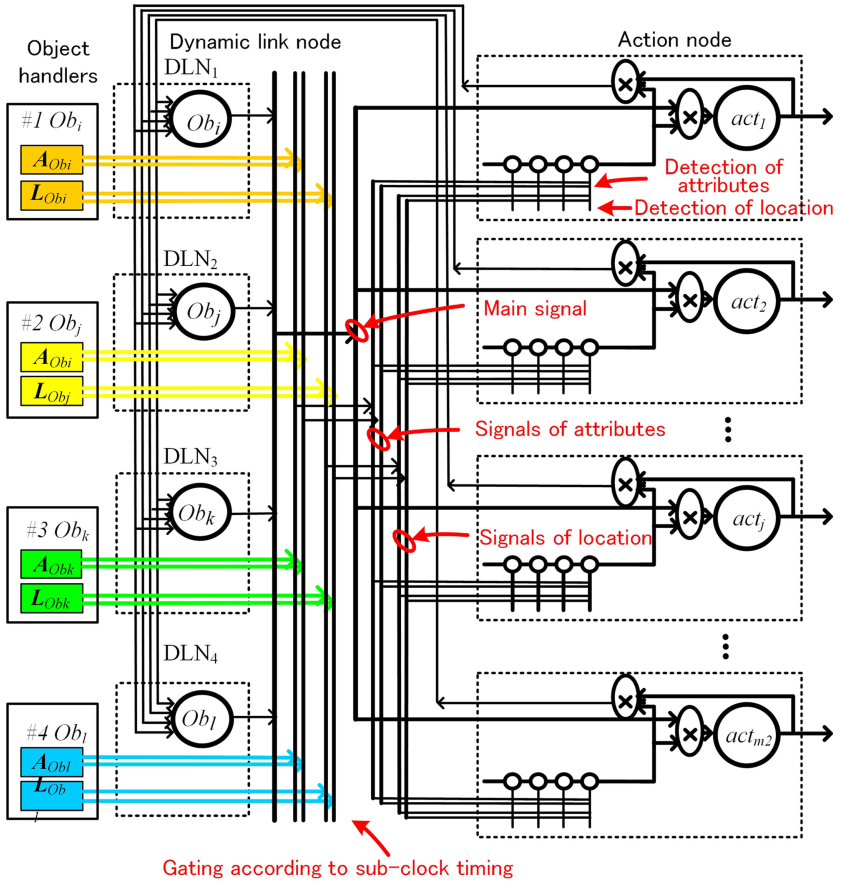


FIGURE 8 | Action decision network using dynamic link node with time division control.

operations for the system itself to change its configuration for adaptation. For (a), the system issued an action instruction at the end of the decision phase so as to perform the action instruction at the fastest and highest priority in line to the basic condition (ii). However, for (b), it is necessary to evaluate the result of the action based on the current system at the relevant time and to instruct the related units in the system to make changes based on the evaluation. It should be noted here that a large number of function units operating in parallel in the system may cause incompatible or inconsistent states among the units.

To deal with these problems, in the postprocessing phase, first, the main states in the system are reconfigured and coordinated. The system updates the information of each object-handler to the latest one, and integrates the same object-handler as the same object when the positions overlap even if they are different object-handler. Through these processes, each object-handler has the latest information of the allocated object. Then information expressing the object's figure with shape and color are reconfigured on the real-image-screen using the object's attributes and location. These attributes and location maintaining by the object-handler were recognized results in the preprocessing phase and were effective for action decision. We call reconfigured information corresponding to a real object existing in environment at that time as a "real-image." The "real-image-screen" is a kind of short-term memory, which maintains the real-images resembling real figures of objects. The reconfiguration of the real-image is performed by reverse processing of the recognition unit using the attribute maintained by the object-handler.

Almost at the same time, processing for two kinds of learning is performed. One is a learning of the recognition unit performed locally, and the other is a learning in relation to action decisions performed as a whole system. The former learning of the recognition unit is performed as the same process when the recognition units execute the reconfiguration of a real image. During reconfiguration, the autoencoder in the recognition unit in **Figure 5** compares decoder's outputs with external stimuli of the object using the comparator and executes self-learning to reduce the difference. Although the real-image-screen is drawn with the output signals of autoencoder, as each signal is checked with each real external stimulus, a highly accurate figure with shape and color can be drawn. Since the recognition units keep learning and correction in each cycle, even if the figure of the object changes slowly over time, it can be recognized as the same object. We presume that the contents of the real-image-screen correspond to what we are aware of when we are looking at things outside in daily life (Meyer and Damasio, 2009).

On the other hand, the latter, learning of action decision is executed as reinforcement learning executed in cooperation with the episodic memory. In the postprocessing phase, the system only writes information for learning into the episodic memory.

This information is read later in the sleeping mode and used for learning of action decision module. In the sleeping mode, the robot is powered on, but it does not respond to external stimuli. Details are shown in the next section.

Transmission of System-Level-Shared-Information and Writing of Learning Information to the Episode Memory

In the latter part of the postprocessing phase, the system makes the state in the system consistent and compatible by widely transmitting and processing the major information for the efficient and speedy next cycle operation. At the same time, information for the system to learn in sleeping mode is written to episodic memory. We focus on the followings as the major information in the system and call this information "system-level-shared-information."

a. The real-image, reconfigured information of the object on the real-image-screen.
b. The information of the evaluated value by the evaluation unit (pleasant/unpleasant).

The system widely transmits this information into the system and processes as follows:

(i) The information of the evaluated value and the object on the real-image-screen is sent into the system *via* the bus.
(ii) The recognition unit that receives the object information from the real-image-screen executes forward recognition processing for the object information. The recognized results, composed of attribute of the object, are transmitted into the system *via* the bus. (For example, if the red circle is on the real-image-screen as an object, "red" and "circle" attribute nodes are output by recognition unit and these attributes are transmitted *via* the bus.)
(iii) The content of object-handler is concurrently updated based on the information from the recognition unit.

In these operations, the reference node corresponding to the meaning of each signal is also excited. As a result, based on the transmission and processing using the system-level-shared-information, the state of each unit connected to the buses and the reference nodes, which are provided parallel in the system, are set in a consistent and compatible state. On these consistent and compatible states, next cycle operation of the parallel units can be executed efficiently and speedy related to basic condition (ii).

The episode memory is connected to the main buses, such as buses related to attributes, action, etc., as shown in **Figure 3B**, and forms a record by collecting information of these buses. The record mainly consists of information reorganized on the bus based on the system-level-shared-information and action instructions. Writing to the episodic memory is executed at the end of the postprocessing phase.

LEARNING PROCESS FOR ACTION DECISION IN BASIC-SYSTEM

This section describes the learning process in the basic-system of the robot.

Execution of Reinforcement Learning

The basic behavior of the robot consists of repeated processes of object search and reward acquisition. The robot walks randomly when there is no object in sight. When one or more objects are captured, the robot selects one preferred object and acts for it as mentioned previously. We call the object selected as desirable object–action pair in action decision phase hereinafter as "target." The target corresponds to the object selected by the robot as an action target or objective. As shown in **Figure 9**, learning for action decision of the robot is performed as a reinforcement learning based on an actor–critic method. The action decision module selects an action as the actor, and the evaluation unit evaluates the action as the critic.

A chain of actions starting from selecting a target object to receiving a reward is taken as one learning episode to which reinforcement learning is performed. This chain of actions is hereinafter referred to as an "event." The term "event" corresponds to the term "episode" commonly used in reinforcement learning, but to avoid confusion with the episodic memory, this article will use the term "event." The robot can handle multiple objects simultaneously, but for simplicity the robot can select up to one target at a time.

When the object is selected as a target, the evaluation unit calculates the value $E^{t}(\boldsymbol{A}^{*t}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t}_{\mathrm{Ob}_i})$ based on the attribute, position of the object by using a value function composed of a neural network. As both the targeted object and not targeted objects are affect the action decisions, even after an object is selected as the target, the robot is not necessarily bound by the targeted object until the reward is received. If more attractive or dangerous objects appear, the robot may change the target to deal with the new object. When the target is switched, the robot starts learning as a different event.

When the robot selects an action for the targeted object, reinforcement learning is performed based on the value$E^{t}(\boldsymbol{A}^{*t}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t}_{\mathrm{Ob}_i})$ as follows:

$$\Delta E(t) = E^{t-1}\left(\boldsymbol{A}^{*t}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t}_{\mathrm{Ob}_i}\right) + R_{\mathrm{real}}(t) - E^{t-1}\left(\boldsymbol{A}^{*t-1}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t-1}_{\mathrm{Ob}_i}\right), \quad (7)$$

$$E^{t}\left(\boldsymbol{A}^{*t-1}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t-1}_{\mathrm{Ob}_i}\right) = \alpha \Delta E(t) + E^{t-1}\left(\boldsymbol{A}^{*t-1}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t-1}_{\mathrm{Ob}_i}\right). \quad (8)$$

Here, $\boldsymbol{A}^{*t}_{\mathrm{Ob}_i}$ and $\boldsymbol{L}^{*t}_{\mathrm{Ob}_i}$ indicate the attribute and the position of the selected object, and $E^{t}(\boldsymbol{A}^{*t}_{\mathrm{Ob}_i}, \boldsymbol{L}^{*t}_{\mathrm{Ob}_i})$ indicates the evaluate value of the selected object at t. $R_{\mathrm{real}}(t)$ indicates the real reward at t. $\Delta E(t)$ in Eq. 7 shows the prediction error in temporal difference learning at t. Based on this prediction error, the critic function performs

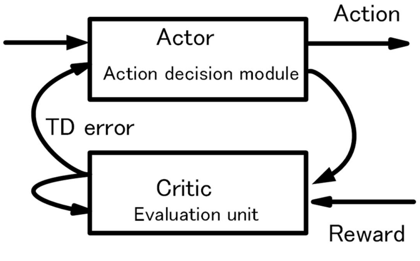


FIGURE 9 | Actor–critic method in the basic-system.

learning as a neural network in the postprocessing phase using learning coefficient α, as shown in Eq. 8. If the prediction error $\Delta E(t)$ is positive, it corresponds to pleasant state or satisfaction with a reward above expectation, and in the negative case, unpleasant state or disappointment with less than expected reward. The above is a case where the robot does not change targets. However, if an object other than the target is selected, it is regarded that the event has been interrupted and the processes in Eqs 7 and 8 are not performed. When the target is switched, the robot starts learning as a different event.

Learning in Cooperation with the Episodic Memory

Learning of the actor composed of a recurrent neural network is not as simple as the critic. The learning is executed in the following two stages using episodic memory, in awake-mode and in sleeping mode. (In the awake-mode, the robot is powered on and can react to external stimuli.)

Writing to the Episodic Memory in Awake-Mode

During awake-mode, the system writes a set of information (referred to as records) related to learning to the episodic memory during each postprocessing phase. The content of the record is composed of the position, attribute, action, output of the value function, and prediction error of the selected object. A sequential chain of records is recorded as a single "event" in the episodic memory. Later, reading the records is done sequentially.

Learning of Recurrent Neural Network in Sleeping Mode

In the sleeping mode, the system reads records from the episodic memory, and learning of the recurrent neural network as the actor is executed using the contents of the records as follows:

(i) The system preferentially selects an event including records with a relatively large prediction error and sequentially reads the records in the event.

(ii) The system changes the coefficient $b_{ij}^{t}(A_{\mathrm{Ob}_i}, L_{\mathrm{Ob}_i}, \mathrm{act}_j)$ of the pattern detector for each record based on the following formula calculated by the information on the record

$$b_{ij}^{t}\left(A_{\mathrm{Ob}_i}^{*t-1}, L_{\mathrm{Ob}_i}^{*t-1}, \mathrm{act}_j^{*t-1}\right) = \beta\Delta E(t) + b_{ij}^{t-1}\left(A_{\mathrm{Ob}_i}^{*t-1}, L_{\mathrm{Ob}_i}^{*t-1}, \mathrm{act}_j^{*t-1}\right), \qquad (9)$$

where β is a learning coefficient. Here, only the part of the recurrent neural network related to the above equation is activated, and the coefficient $b_{ij}^{t}(A_{\mathrm{Ob}_i}, L_{\mathrm{Ob}_i}, \mathrm{act}_j)$ is changed in the direction along ΔE.

The reasons that the learning of the actor using episodic memory is performed during sleeping mode are as follows:

a. To execute the learning shown in Eq. 9, it is necessary to activate only the part of the recurrent neural network related to learning. Other parts of network cannot operate at the same time. If the recurrent neural network learns during awake-mode, the network must temporarily stop responding to external stimulus during the learning process. The robot operation will have to stop intermittently during learning. Assuming the robot was an animal, it will not be able to react to dangerous conditions quickly if it tried learning while it was awake.

b. Utilizing learning information after recording in episodic memory has some advantages. One is that the system can learn efficiently by utilizing experiences, based on selection, or repeating large impact events by looking back on past experiences. The other is the system enables relatively stable adaptation with less risk of over-training by not learning immediately when an event occurs.

In the case of an animal, execution of the learning in sleeping mode causes the animal to be in relative risk against predators during sleep, but overall there is merit for the animal to learn during sleep.

CONSCIOUSNESS IN THE BASIC-SYSTEM

Basic Hypothesis on Consciousness on the Basic-System

Consider the system-level-shared-information shown in the basic-system from the viewpoint of animals. We presume that an animal's brain is composed of (a) functions that respond automatically or semi-automatically according to stimuli and (b) functions for system-level processing such as action decisions. The automatic or semi-automatic functions operate in parallel under loose coordination.

When an animal acts as one individual or one system, such as when going toward a prey or escaping from a predator, it is necessary for these functional units in the brain to have tightly related cooperation based on system-level decisions. For this purpose, it is an effective way to share consistent and clear information of objects and directions of action, such as approach or avoidance among functional units which should be tightly related for cooperation at the time. Based on this shared information, each functional unit performs consistent simultaneous operation so that the animal's ability can be demonstrated as much as possible. In particular, "pleasant/unpleasant" is basic information that indicates either the necessity of action as individuals, approach or avoidance, and needs to be notified as quickly as possible. By using this pleasant/unpleasant information and object information in combination, to move more closely to prey or avoid predator becomes possible for the brain.

A unicellular paramecium backs away when it hits an obstacle ahead and swims at a speed that is more than twice the usual against a stimulus from behind. At that time, Paramecium sends information concurrently to thousands of cilia of Paramecium, organs of for move, by changing in membrane potential or ion concentration, in accordance to the stimulus received by the sensor. With this information, a large number of cilia perform a consistent operation along the direction of movement of the paramecium, as one individual (Kutomi and Hori, 2014). This indicates that even if the organism is extremely simple, if it is composed of many functional units and prompt action is required, to send system-level-shared-information to related organs all at once is necessary.

Although the contents of our "awareness" at each moment are diverse and contain various subtle elements, one of the main contents of awareness is the perceived world around us composed of objects, and the feeling of our "self" in this perceived world. Based on the above, we hypothesize that the main part of what we recognize as phenomenal consciousness corresponds to system-level-shared-information in the basic-system. We assume that even primitive animals have "awareness," to adapt autonomously as a single system, and execute information transmission and processing corresponding to system-level-shared-information.

In addition, since an animal mainly acts using automatic or semi-automatic functions, system-level-shared-information is issued only when an action decision as a whole system is needed. If functions that are automatically or semi-automatically operated in parallel can respond appropriately to stimuli, system-level-shared-information is not issued. When we ride a bicycle for the first time, we are initially aware of the operations required to ride a bicycle, including pedaling, steering, and balancing. But when we get used to riding a bicycle, we are not aware anymore of the individual operations. Initially, the bicycle riding operations become the objective of the system-level action decision. As the semi-automatic processing function begins to work, the necessity to operate a bicycle disappears at the system-level. At this time, the system-level-shared-information for riding a bicycle is not required and is not generated anymore.

Logical Organization of Consciousness and Self

Figure 10 shows the perceived space logically composed of the system-level-shared-information. In this space, the state of the evaluation unit and objects are the main elements. An evaluation unit located at the origin of the space evaluates the object. The relationship between the object and the robot including the system is the basis of the operation of the autonomous adaptation. Each object had been treated as a bundle of attributes etc. as we have mentioned. However, since the robot itself is composed of a large number of entities, the relationship between the robot and the object cannot be briefly expressed unless the robot is bundled too, or represented by something.

Since the robot operates on a complex interaction of various motor and functional units, bundling of some specific entities is not appropriate. Between the robot, and the object, "what kind of action the robot is going to do with respect to the object" is

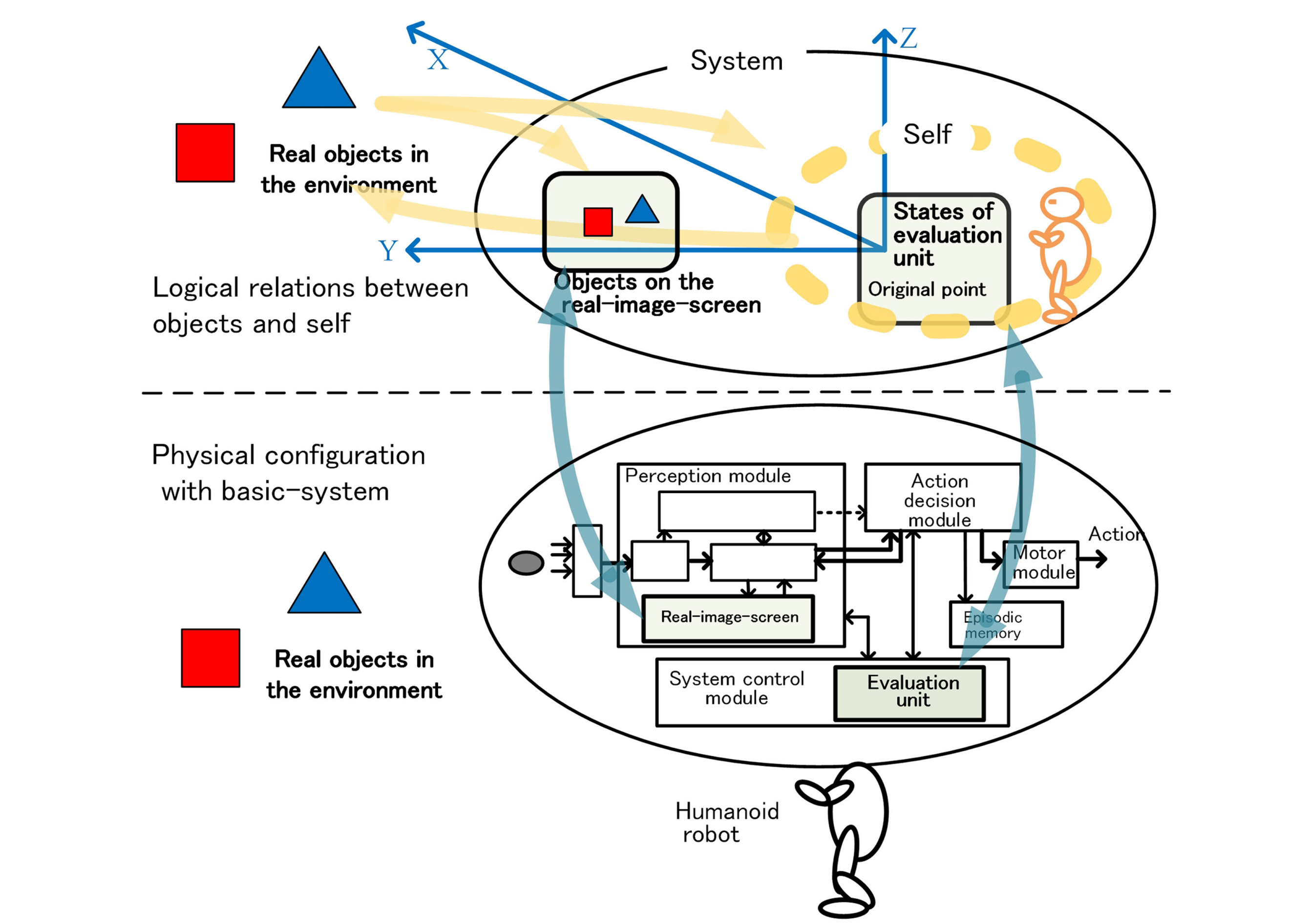

X
System
Z
Self
Real objects in the environment
States of evaluation unit
Original point
Y
Objects on the real-image-screen
Logical relations between objects and self
Physical configuration with basic-system
Perception module
Action decision module
Action
Motor module
Real-image-screen
Episodic memory
System control module
Evaluation unit
Real objects in the environment
Humanoid robot

FIGURE 10 | Logical relations between objects and self on the basis of the physical configuration.

important, and "what is the bundled unit as the entity of the robot" is not necessarily required. From this point of view, the state of the evaluation unit briefly and basically shows the direction of action decision as a robot to the object, and implicitly represents the robot including the system itself.

Based on this view, we regard the state of the evaluation unit as a kind of "self" as in the bundle theory of self by Hume (Pike, 1967; Smith, 2017). The "self" existing at the origin forms the relationship between "self" and the objects. The "self" sees and copes with the objects. We speculate that this relationship contributes to the awareness of the first-person perspective as if the homunculus in our brain sees the outside world (the orange robot in **Figure 10**).

The Binding Problem and the Delay Time of Libet's Experiment

Based on the above hypothesis, the Binding problem and the delay time of Libet's experiment can be accounted for as follows. An outline is shown in **Figure 11**.

Binding Problem

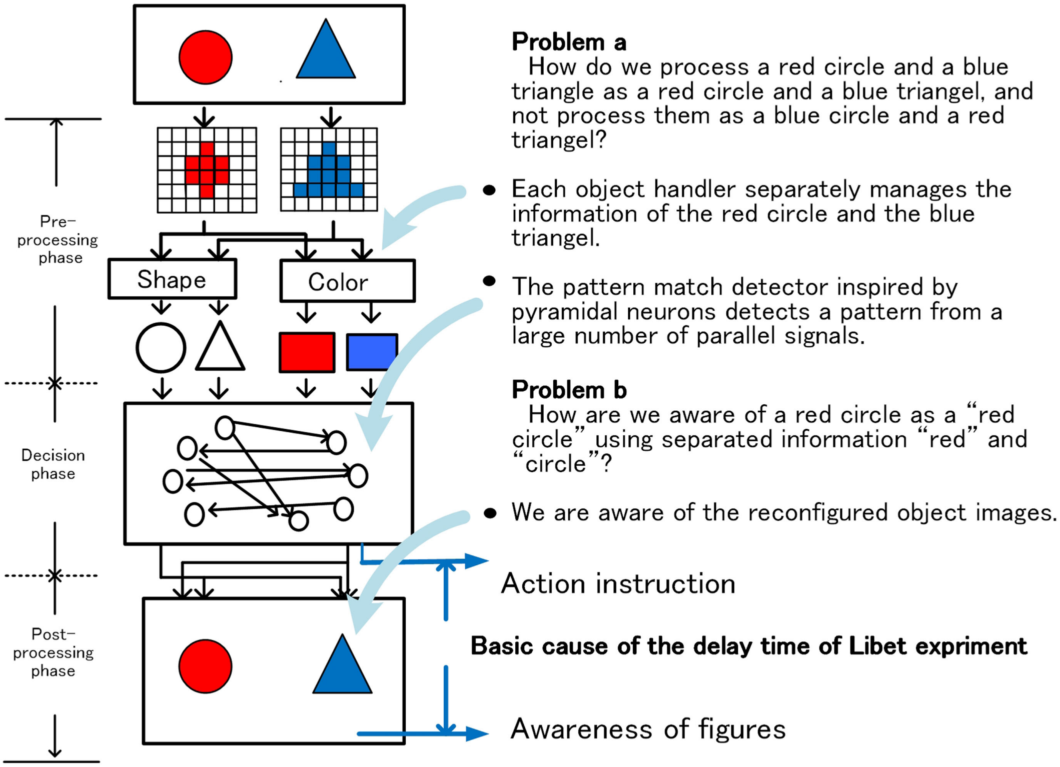
As **Figure 11** shows, the brain is known to process shapes and colors with different functional units. In the case of a red circle and a blue triangle, shape and color are processed as separate signals by separate functional units. In relation to this, there is an unsolved problem, known as the Binding problem, in the brain (Kahneman and Treisman, 1992; Pylyshyn, 2001; Meyer and Damasio, 2009; Xu and Chun, 2009; Bays et al., 2011; Feldman, 2013). The binding problem is roughly expressed as the following two problems.

Problem a. How do we process a red circle and a blue triangle as a red circle and a blue triangle, and not process them as a blue circle and a red triangle?

Problem b. How are we aware of a red circle as a "red circle" using separated information "red" and "circle"?

Based on the hypothesis that the main contents of awareness correspond to system-level-shared-information, the system provides the answer to the problems as follows.

In the system, each circle and triangle is allocated to different object-handlers as different objects and managed. The object-handler instructs the related functional units to recognize (pattern recognition, color recognition, etc.) the allocated object and maintains the resultant signal as a set of parallel signals composed of shape and color. Information on the shape and color of the red circles keeps held by the object-handler until the object disappears. Information of each object is input to the action decision module as a set of parallel signals under the control of the allocated object-handler. In the action decision module, there are many action nodes corresponding to the type of various actions. And each action node has a lot of detectors that detect matched parallel signal pattern from the thousands of parallel signals. Using this function, each action node detects only the signal pattern that the corresponding action is deemed necessary from the thousands of parallel signals, and reacts to the signal pattern. This means that, in the case of animals that eat, for example, red apples but not blue prunes, the node for eat has a detector that detects "red" and "apple." That is, although information of the shape and color of an object are processed separately, the object-handler manages it



Problem a
How do we process a red circle and a blue triangle as a red circle and a blue triangel, and not process them as a blue circle and a red triangel?

- Each object handler separately manages the information of the red circle and the blue triangel.
- The pattern match detector inspired by pyramidal neurons detects a pattern from a large number of parallel signals.

Problem b
How are we aware of a red circle as a "red circle" using separated information "red" and "circle"?

- We are aware of the reconfigured object images.

Action instruction

Basic cause of the delay time of Libet expriment

Awareness of figures

FIGURE 11 | The binding problem and basic cause of the delay time of Libet's experiment.

as a parallel signal belonging to the same object. Furthermore, a large number of parallel signals are directly checked for action decision while being parallelized using the mechanism inspired by pyramidal cells in the cerebrum. This makes it possible to explain the Problem a.

Furthermore, after deciding the action, the recognition system operates in the reverse direction to reconfigure the "red circle" on the real-image-screen, using the information, "red" and "circle" in the allocated object-handler. Then, we are aware of the "red circle" on the real-image-screen as a part of system-level-shared-information. In this way, the Problem b can be explained.

When combining and processing parallelized signals, the timing adjustment between signals is required. Without timing adjustment, it is not possible to perform processing based on the mutual relationship between signals appropriately. In general, when the number of stages of timing adjustment increases, the response time of the system becomes long because it is necessary to wait the signal arriving at the latest and to spend time to process signals at each processing stage. From this point of view and basic condition (ii), we speculate that the system uses a method in which the number of stages of timing adjustment relating to slow response is minimized.

The Time Delay in Libet's Experiment

Famous experiment of Libet shows that our intentional movements are initiated before we become conscious to act, and have been calling a lot of debate so far (Libet, 2004). As we have repeatedly mentioned, in the system, since high priority is given to quick responses, action instructions are issued immediately after the decision phase and "awareness" occurs late in the postprocessing phase. The time difference shown in **Figure 11** does not accurately correspond to the delay time indicated by the experiment of Libet, but we consider that it shows a basic cause of the delay time. From this viewpoint, we consider that our hypothesis for phenomenal consciousness is consistent with the Libet's experiment.

PROPOSAL OF THE EXTENDED-SYSTEM

We have proposed the basic-system as an autonomous adaptive system that performs habitual behavior. In relation to consciousness, we have shown that awareness is an important operation for executing parallel processing. However, the basic-system cannot perform "goal-directed behavior," consisting of setting a goal and conducting actions to achieve that goal through various attempts. Also functions that manipulate recollected objects, which are an important element of our conscious experience, are not incorporated. To realize these functions, we propose the "extended-system" as an extension of the basic-system.

Outline of Goal-Directed Behavior in the Extended-System

Action Suspending

In the basic-system, an action instruction selected for the object is immediately executed in the decision phase, and the evaluation process for the action is executed in the postprocessing phase. In the extended-system, if necessary, the action on the object is suspended and no action is taken. For example, in cycle t, without taking action, the system predicts the reward of action for the object. Then, in cycle $t + 1$, the system can decide an action considering the predicted reward. This example shows that if the system temporarily suspends an action instruction, adaptive action considering multiple cycles becomes possible. We presume that this suspension of an action is related to "*the ability to delay immediate gratification for the sake of future consequences*" of children in the marshmallow test in psychology (Mischel, 2014).

Fast Decision and Slow Decision

In the extended-system, an action decision aiming for quick response (fast decision) and an action decision aiming at a higher level of adaptation with slow response speed (slow decision) are used depending on the situation. In the fast decision, the basic function corresponding to the basic-system operates with quick response, based on reinforcement learning. On the other hand, in the slow decision, the extended-system takes the risk of putting real actions on hold, allowing the system to aim for a higher level of reward.

When operating in slow decision, more processing time and resources in the system are used than in fast decision. We surmise that in the slow decision, the system needs some kind of "motivation" to take risks, use higher resources, and to try to achieve higher rewards. In the extended-system, in addition to the pleasant/unpleasant state of the basic-system, a value corresponding to motivation is maintained and managed as an indicator (degree of motivation), and execution of slow decision is controlled according to this value.

Assumed Primitive Behaviors

The goal-directed behavior consists of a chain of slow decisions aimed at achieving the goal. A primitive example of a series of decisions from detection of objects to acquisition of reward through various actions is shown in **Figure 12**.

Configuration and Functions of the Extended-System

To configure the extended-system as simple as possible, the following policies were adopted:

a. The extended-system is constructed using the functions of the basic-system as much as possible. Additional functions are minimized.
b. The function to be configured as a new circuit is limited to functions commonly used, or functions requiring high speed.
c. Information for high-level or detailed behavior is stored in long-term memory as much as possible and read out as necessary.

Based on these policies, when a goal-directed behavior is performed, although the number of times of reference to long-term memory and response time increases, it is possible to achieve a sophisticated adaptation at low cost. In addition, the time required for learning can be shortened as compared with the case of using a dedicated neural network circuit. This is because it is possible to record to long-term memory in a short time as compared with learning time of the dedicated neural network circuit.

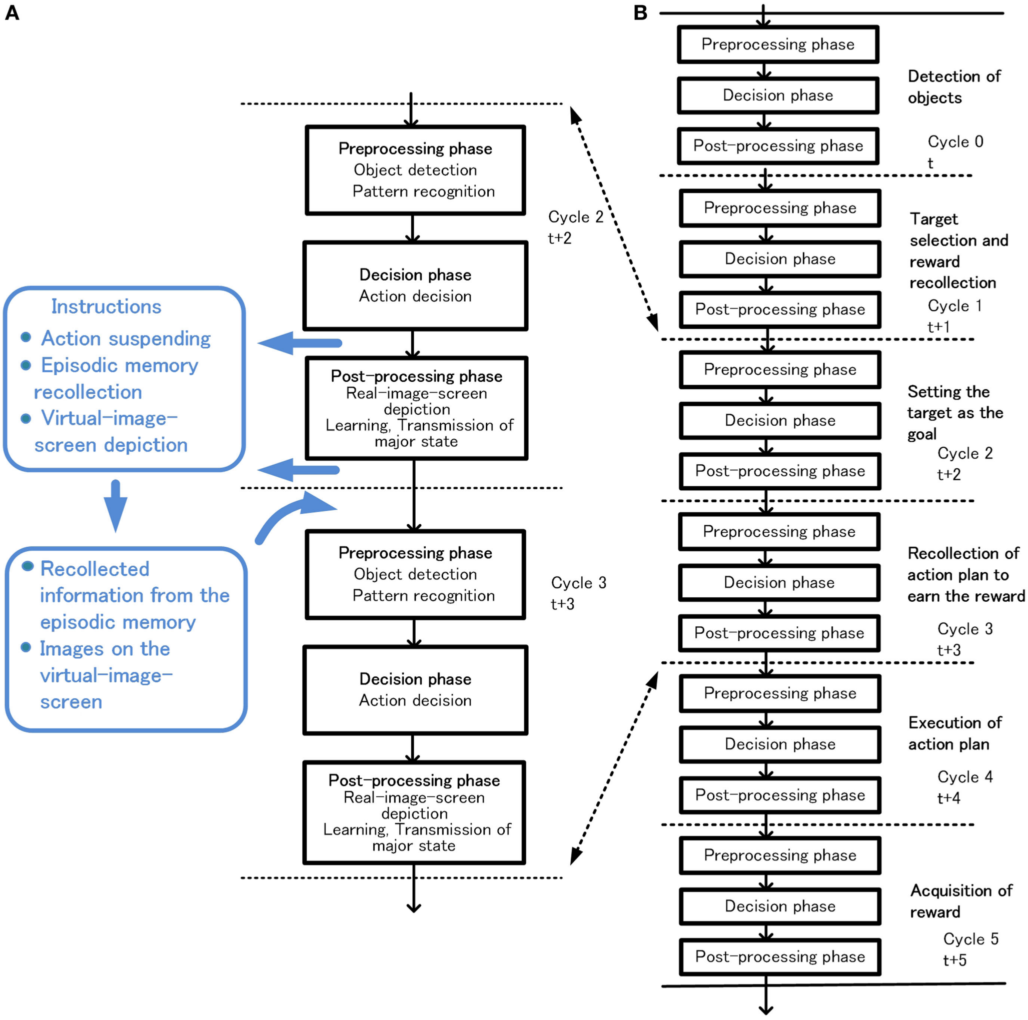


FIGURE 12 | Outline of behavior in the extended-system. Coordination between cycle 2 and cycle 3 utilizing action suspending, episodic memory recollection or virtual-image-screen depiction **(A)**, and an example of goal-directed behavior **(B)**.

Under the above policies, the following dedicated functions were provided. Outline of functional extensions in the extended-system is shown in **Figure 13**. The bus configuration of the extended-system is shown in **Figure 14**. The orange box in the **Figure 14** shows the main unit added to the basic-system.

Extension of Basic Functions in the Basic-System

The following functions are expanded in the extended-system.

Extension of Object Handling Function

In the extended-system, the goal is expressed by three elements, (1) what is targeted, (2) what actions to be taken on that object, and (3) what can be earned as reward. A single object-handler can hold this set of object, action, and evaluation value to express a goal. It is not necessary that all three elements of object, action, evaluation value is available.

Extension of Action Decision-Related Functions

The output of the action decision module was only an action instruction in the basic-system, but in the extended-system, instructions for suspending action, setting as a goal, recalling of long-term memory, and handling images as an object are added.

Addition of Image Manipulation Function

Functions related to the manipulating image, information generated inside the system, are added as common functions.

Buffer Memory for Expressing Patterns of Images (Virtual-Image-Screen)

To manipulate information generated within the system, such as recollected objects, in the same way as information of real objects existing actually at that time, a temporary buffer memory for expressing patterns of images, which we named "virtual-image-screen," is provided. We call reconfigured information not corresponding to a real object existing in the environment at that time as a "virtual-image." The "virtual-image-screen" is a kind of short-term memory, which maintains the virtual-images. Recollected contents from the long-term memory are depicted in the virtual-image-screen when in the awake-mode. Object-handlers can capture objects in the virtual-image-screen similarly to capturing

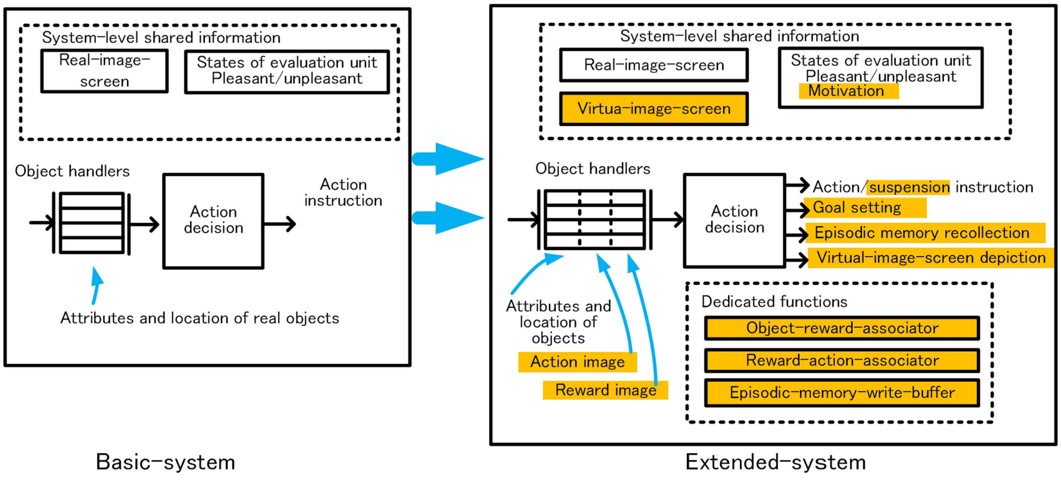

System–level shared information
Real–image–screen
States of evaluation unit Pleasant/unpleasant
Object handlers
Action decision
Action instruction
Attributes and location of real objects
Basic–system
System–level shared information
Real–image–screen
Virtua–image–screen
States of evaluation unit Pleasant/unpleasant
Motivation
Object handlers
Action decision
Action/suspension instruction
Goal setting
Episodic memory recollection
Virtual–image–screen depiction
Attributes and location of objects
Action image
Reward image
Dedicated functions
Object–reward–associator
Reward–action–associator
Episodic–memory–write–buffer
Extended–system

FIGURE 13 | Outline of functional extensions in the extended-system.

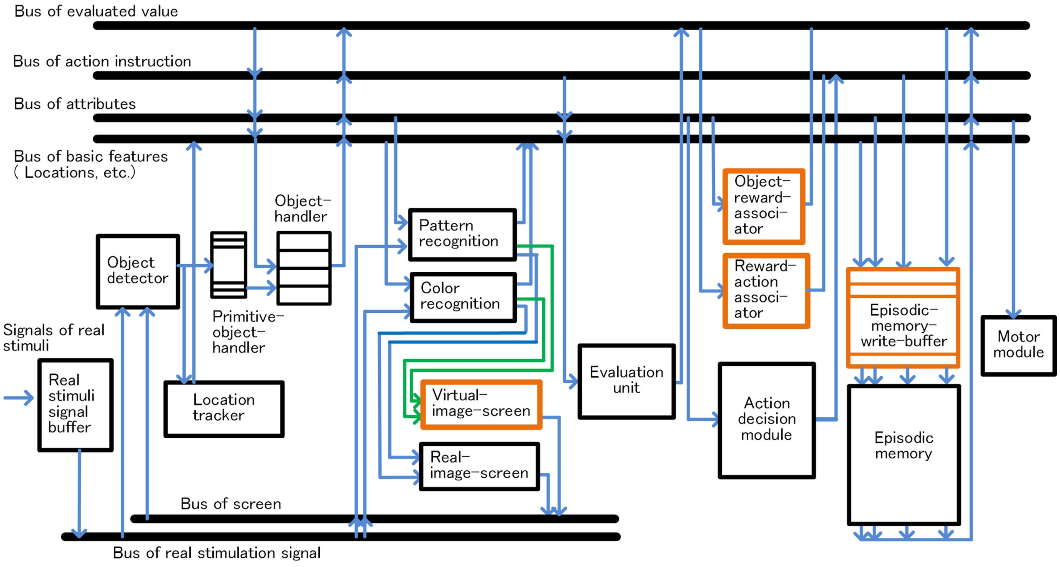


FIGURE 14 | Bus configuration of the extended-system.

objects in the real-image-screen, so objects captured from the virtual-image-screen can influence action decisions same as real objects in the basic-system. Based on this method, the extended-system can decide actions using past experiences or knowledge in the awake-mode.

The virtual-image-screen corresponds to our mental imagery as shown below. We use "mental imagery" as defined by Kosslyn (1994).

a. In the postprocessing phase, objects on the virtual-image-screen are reconfigured using the attribute of objects same as with the objects on the real-image-screen. The contents of the virtual-image-screen are subject to object detection like the real-image-screen. In addition, the contents of the virtual-image-screen are transmitted to a wide range of the system as a component of system-level-shared-information.

b. The signals for expressing the virtual-image-screen which are output from the recognition units (green lines in **Figure 14**), are not compared with the real stimulus by the autoencoder. In the case of the real-image-screen, comparison with real stimulus is executed by the autoencoder, so the system can express images on the real-image-screen clearly. In the case of the virtual-image-screen, the reconfigured images are blurred because there is no comparison with real stimulus. The extended-system uses a bus in which meaning is managed by the reference node in common with external stimuli. Thus, internally generated contents can have the same meaning corresponding to external stimuli.

By this configuration, the extended-system can treat the clear contents corresponding to real things in the real-image-screen, and the blurred contents generated inside the system in the virtual-image-screen concurrently. These contents correspond to how the human brain is aware, using a clear image of the real world and a blurred mental imagery.

Action Control by Action Patterns

Recollecting and execution of action in the extended-system is realized as a function to connect action instruction signals with visual action patterns. Visual action patterns are represented as a kind of image in the virtual-image-screen and manipulated as an object. When the system outputs an action pattern on the virtual-image-screen, the connected action signal is excited, and the action decision unit selects the corresponding action with highest priority. This means that the extended-system can be directed to perform that action by outputting a certain action pattern on the virtual-image-screen. In addition, the system can deliver the visual action pattern to the next cycle as part of the system-level-shared-information. In this manner, actions are treated as objects represented by a kind of visual action patterns. This function was adopted on the basis of findings of the mirror neuron (Rizzolatti et al., 2014).

The State of Evaluation Unit for Manipulating Reward

For the extended-system to handle reward as a goal, it is necessary to manipulate the state, such as pleasant or unpleasant, as a kind of object or signal independent of the system's own evaluation unit state. The evaluation unit of the extended-system can have the following two states at the same time.

Effective-Excitation (EE) State. Reference nodes corresponding to the state are activated, and the activation is transmitted to the whole system. System-level learning is executed based on this state.

Non-Effective-Excitation (non-EE) State. Reference nodes corresponding to the state are activated, but the activation is not transmitted to the whole system and effective only as signals representing information of the state. The system-level learning is not executed based on the state. Signals of non-EE state are used for handling reward such as goals.

Object–Reward and Reward–Action Associating Function

Dedicated circuits, episodic memory write buffer, object–reward-associator, and reward–action-associator are provided to execute reward prediction from the target object and desirable action recollection from reward. The episodic memory write buffer maintains recent results of the action decision module for a 100 cycles before storing the episodic memory. The object–reward-associator, consisting of a bidirectional pair of neural networks, associates a target object and a reward value. Likewise, the reward–action-associator, consisting of a bidirectional pair of neural networks, associates a reward and an action.

The learning of object–reward-associator is performed by simultaneously exciting a target object and reward information on episodic memory write buffer in pairs, and supplying pairs of inputs and outputs to the unit through the bus. The neural network modifies the weight so that the supplied signal pair is associated with each other. In the reward–action-associator, learning is executed in the same way. These learning operations are executed under time shared control within the postprocessing phase.

Extended-System Operation

Table 1 shows an example of robot operations with the extended-system for a goal-directed action, including the changes of state in system-level-shared-information corresponding to the robot awareness. This example was modeled with reference to the goal-directed behavior experiment using monkeys by Matsumoto et al. (2003) and the Experimental Cognitive Robot by Haikonen (2012).

The robot determines actions based on the conscious information in the previous cycle and summarizes the result to the next conscious information. This process is repeated. Defining that the perceived world for the robot is the world of what the robot is aware or conscious of, from the viewpoint of the robot, the robot decides and acts on the world using summarized conscious information. We think that this flow of state changes in the system-level-information corresponds to the flow of consciousness for us humans.

In **Table 1**, it was assumed that the motivation level of the robot is sufficiently high for performing the slow decision. If the robot is exhausted and the motivation is lowered, the robot ignores the detected objects. When the robot receives rewards, the system evaluates the reward as pleasant/unpleasant according to the difference between the expected value and the obtained value. Primitive learning is performed by reflecting this value (pleasant/unpleasant) in the learning of the object–reward-associator and the reward–action-associator. However, learning methods throughout the entire robot including motivation adjustment are still under consideration.

ON CONSCIOUSNESS

Consciousness as Awareness

We assumed that the brain basically works like the basic-system in principle because the brain should perform at its full potential as a parallel-processing system. In this case, the brain selects and decides the fastest and most efficient action, and responds immediately. After the action decision, postprocessing is done throughout the brain and prepared for the next stimulus.

In this postprocessing, scattered information is organized/integrated, learning based on reward is executed, and these results are notified through the brain. The phenomenon of awareness corresponds to the most important notified information that is "system-level-shared-information," composed of states of evaluation unit and objects. This information forms a space where the state of evaluation unit is located at the origin and various objects exist together. This space corresponds to the "subjective space" that we are aware of on a daily basis, and the evaluation state corresponds to "self."

One of the characteristics of phenomenal consciousness is "integration of information." Tononi (2012) explained in the integrated information theory using Φ, but we consider that Φ is unnecessary for explanation of consciousness. Through the

TABLE 1 | An example of robot operations as a minimal level goal-directed action.

Operations of the robot	**Operations in the system**	**Awareness of the robot system-level-shared information**			
		Real-image-screen (real-IS), virtual-image-screen (virtual-IS)		**Evaluated value Effective excitation (EE), non-effective excitation (non-EE)**	
		Real-IS	**Virtual-IS**	**EE**	**Non-EE**
1. Searching objects The robot detects nothing in the environment, then walks randomly to detect objects	a. Walking randomly for searching objects is installed as a basic function				
2. Detection of objects The robot detects a **red box** and a **blue box** as object, then stops walking	a. Two object-handlers maintain information of the **red box** and **blue box**, respectively, and these are recognized b. Reconfigured information of these boxes is depicted on the real-image screen	**Red box** **Blue box**			
3. Target selection and reward recollection The robot selects the **blue box** as a target and recollects reward related to the target	a. The action decision module selects the **blue box** as a target, then object–reward-associator outputs reward recollection related to the **blue box** b. State of the evaluation unit becomes **Pleasant** in **non-EE** c. No actual action instruction	**Red box** **Blue box**			**Pleasant**
4. Setting the target with reward as the goal The robot sets the **blue box** including reward as the goal	a. The object-handler allocated to the **blue box** maintains information of reward as a goal b. No actual action instruction	**Blue box**			**Pleasant**
5. Recollection of action to earn the reward	a. The behavior of the action, output of the reward–action-associator, are depicted in the virtual-image-screen b. No actual action instruction	**Blue box**	**Touching action**		**Pleasant**
6. Execution of action plan The robot touches the **blue box**	a. The robot is charged by touching the **blue box** b. The evaluation unit becomes **Pleasant** in **EE**	**Blue box**		**Pleasant**	
7. Acquisition of reward The robot is **Pleasant** by charge really		**Blue box**		**Pleasant**	
8. Execution of learning through these experiences	a. Learning is executed in the sleeping mode	–	–	–	–

Significant states and objects are depicted in color or bold (for highlighting purposes).

processing of the decision phase and postprocessing phase shown so far, "integration of information" as a phenomenon can be generated. Based on our model, we can explain the binding problem and show the basic causes of delay in Libet's experiment, which indicates that Φ is unnecessary. Consciousness is a necessary function for the brain to perform at the full potential as a central control system of an animal.

Our proposed system is close to Haikonen's robot and Franklin's system (Haikonen, 2012; Franklin et al., 2014), and to proceed in the future, it is necessary to incorporate the various functions proposed in these systems. However, our proposed system is different from their methods in the core design regarding action decision and consciousness. In GWT, dedicated processors compete for the right in the limited storage area called Global Work Space, and the action plan of the processor that got this right is broadcasted and conscious. In HCA, dedicated processors attempt to communicate with each other, and the main successful communication becomes conscious. For the following reasons, our proposal is more appropriate than GWT and HCA.

(i) In GWT and HCA, the information and actions to be selected are determined by mutual relationship among individual dedicated processors. We think that reflections of what is desirable as a system are not sufficient in this selection. In our model, the system chooses the optimum pair based on the desirability as a system from "object and action pair." The system further performs reinforcement learning using episodic memory for individual combinations. In addition, in our model, it is a choice of optimal object and action pair, so multiple pairs cannot be allowed to exist simultaneously. This explicitly explains the "unity" which is the basic characteristic of consciousness (Brook and Raymont, 2017).

(ii) In GWT, it is claimed that broadcasting is the main factor of awareness or consciousness. We presume this broadcast correspond to wide-area transmission in postprocessing in our model. However, in postprocessing, various information that we are not phenomenally conscious of is simultaneously transmitted in a wide area. We assume that we are only aware of system-level-shared-information, not simply the information transmitted over a wide area. The system-level-shared-information is composed of the state of the evaluation unit and the state of the object. We speculate that activation of the evaluation unit that represents "self" is indispensable factor of conscious experience.

(iii) Dehaene and Changeux (2011) assert the validity of GWT based on brain observations such as fMRI, event-related potentials. However, since our model shown in this article is expected to be observed as a phenomenon similar to GWT, it also supports the validity of our model. In our model, the recurrent neural network optimization process in the decision phase roughly corresponds to the activation centered on the frontal lobe, and the processing in the postprocessing phase roughly corresponds to the activation of a wider area including the occipital lobe.

Consciousness as an Important Function for the Complex Brain

In the extended-system, we think that the chain of "conscious information," which directs actions toward a goal with intention, corresponds to our "proactive" conscious state. In addition, it is important that the conscious information can be handled as objects in the next cycle. Since the conscious information expresses the state of the system in a summarized form, the system can decide an action efficiently and easily by using this summarized information. This shows that by manipulating the conscious information, complex systems such as the extended-system can be controlled efficiently and easily. We speculate that it is through this function of consciousness that we can "think" and make decisions without being aware of the complexity of the human brain.

DISCUSSION

Duality Model

Duality models of human behavior, such as fast/slow thinking in the behavioral economics field and impulsive/reflective system in the social psychology field are well known (Deutsch and Strack, 2004; Kahneman, 2011). We predict that this duality arises from fast decision due to direct fast responses, and slow decision due to sophisticated adaptation at the expense of response speed, depending on the circumstances in the extended-system. An action mainly composed of fast decisions appears as a fast or impulsive action, and an action mainly composed of slow decisions appears as a slow or reflective action.

From another point of view, a fast decision shows passive and reactive behavior against the stimulus, such as prompt decision as to whether or not to eat when bait appears. On the other hand, in a slow decision, such as when a stimulus that is not directly related to bait has appeared, shows a proactive behavior that looks ahead toward an intended goal.

Goal-Directed Behavior Incorporating a General-Purpose Computer Like Function

Although there are various ways to perform goal-directed behavior, the main aim of the proposed extended-system was "to realize advanced adaptation at a relative low cost by sacrificing response speed." We assumed that it is important to realize goal-directed behavior through a combination of common or general-purpose circuits together with long-term memory as designed in the early computer EDVAC (von Neumann, 1945). We predict that a conscious autonomously adaptive system that achieve goals set by itself will become a powerful control system for the humanoid robots by incorporating a kind of von Neumann type computer as extended functions.

CONCLUSION

We proposed a basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot. We think resembling the human brain at the level of the basic logical structure, architecture, is a meaningful way of designing a control system for a truly useful humanoid robot. Interaction or communication between humans and humanoid robots will be much easier if both sides shared the same behavior characteristics based on the same architecture, such as consciousness or duality. However, the proposal in this article currently remains at the architecture design level, and verification through simulation is still only partial. We plan to further refine the system configuration with reference to the results of previous research by Haikonen and Franklin et al., as well as new findings. Evaluation of the dynamic characteristics of the system through simulation is also planned.

AUTHOR CONTRIBUTIONS

YK designed the basic architecture of this autonomous adaptive neural network system. YK and KM together discussed and built upon the basic design to produce the complete architecture of the proposed system and together wrote this article.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Pentti O. Haikonen for his valuable advice. The authors also thank Mr. Shoji Inabayashi and the students of Kinouchi Laboratory for their valuable comments during the course of the research.

REFERENCES

Anderson, J. (1983). Cognitive and psychological computation with neural models. *IEEE Trans. Syst. Man Cybernet.* SMC-13, 5. doi:10.1109/TSMC.1983.6313074

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

Baars, B., and Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: global workspace theory and IDA. *Neural Netw.* 20, 955–961. doi:10.1016/j.neunet.2007.09.013

Bays, P., Wu, E., and Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia* 49, 1622–1631. doi:10.1016/j.neuropsychologia.2010.12.023

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi:10.1561/2200000006

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends Cogn. Sci.* 15, 567–575. doi:10.1016/j.tics.2011.11.001

Brook, A., and Raymont, P. (2017). *The Unity of Consciousness*. The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/sum2017/entries/consciousness-unity/ (Accessed: November 13, 2017).

Coward, L. (2013). *Toward a Theoretical Neuroscience: From Cell Chemistry to Cognition*. Dordrecht: Springer.

Dehaene, S. (2014). *Consciousness and the Brain Deciphering how the brain codes our thoughts*. New York: VIKING.

Dehaene, S., and Changeux, J. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 201–227. doi:10.1016/j.neuron.2011.03.018

Dennett, D. (1994). "Consciousness in human and robot minds," in *IIAS Symposium on Cognition, Computation and Consciousness*, Kyoto.

Deutsch, R., and Strack, F. (2004). Reflective and impulsive determinants of social behavior. *Person. Soc. Psychol. Rev.* 8, 220–247. doi:10.1207/s15327957pspr0803_1

Deutsch, R., and Strack, F. (2006). Duality models in social psychology: from dual processes to interacting systems. *Psychol. Inq.* 17, 166–172. doi:10.1207/s15327965pli1703_2

Dezfouli, A., and Balleine, B. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Comput. Biol.* 9:e1003364. doi:10.1371/journal.pcbi.1003364

Feldman, J. (2013). The neural binding problems. *Cogn. Neurodyn.* 7, 1–11. doi:10.1007/s11571-012-9219-8

Franklin, S., Madl, T., D'Mello, S., and Snaider, J. (2014). LIDA: a systems-level architecture for cognition, emotion, and learning. *IEEE Trans. Auton. Mental Dev.* 6, 1. doi:10.1109/TAMD.2013.2277589

Franklin, S., and Patterson, F. G. Jr. (2006). "The LIDA architecture: adding new modes of learning to an intelligent, autonomous, software agent," in *Proceedings (Integrated Design and Process Technology)* (San Diego: Society for Design and Process Science).

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

Golden, M. (1986). The brain-state-in-a-box neural model is a gradient descent algorithm. *J. Math. Psychol.* 30, 73–80. doi:10.1016/0022-2496(86)90043-X

Golden, M. (1993). Stability and optimization analyses of the generalized brain state in a box neural network model. *J. Math. Psychol.* 37, 282–298. doi:10.1006/jmps.1993.1017

Gray, J., and Reuter, A. (1993). *Transaction Processing: Concepts and Techniques*. San Francisco: Morgan kaufmann.

Gremel, C., and Costa, R. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat. Commun.* 4, 2264. doi:10.1038/ncomms3264

Haikonen, P. (2003). *The Cognitive Approach to Conscious Machines*. Exeter: Imprint academics.

Haikonen, P. (2007). *Robot Brains Circuits and Systems for Conscious Machines*. Chichester: John Wiley & Sons.

Haikonen, P. (2012). *Consciousness and Robot Sentience*. Toh Tuck Link: World Scientific.

Hart, G., Leung, B., and Balleine, B. (2013). Dorsal and ventral streams: the distinct role of striatal subregions in the acquisition and performance of goal-directed actions. *Neurobiol. Learn. Mem.* 108, 104–118. doi:10.1016/j.nlm.2013.11.003

Hwang, K., and Briggs, F. (1984). *Computer Architecture and Parallel Processing*. New York: McGrow-Hill.

Jeffers, J., and Grabowski, A. (2017). Individual leg and joint work during sloped walking for people with a transtibial amputation using passive and powered prostheses. *Front. Robot. AI*. 4:72. doi:10.3389/frobt.2017.00072

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books, Pearson.

Kahneman, D., and Treisman, A. (1992). The reviewing of object files: object specific integration of information. *Cogn. Psychol.* 24, 175–219. doi:10.1016/0010-0285(92)90007-O

Kasai, H., Fukuda, M., Watanabe, S., Hayashi-Takagi, A., and Noguchi, J. (2010). Structural dynamics of dendritic spines in memory and cognition. *Trends Neurosci.* 33, 121–129. doi:10.1016/j.tins.2010.01.001

Kinouchi, Y. (2009). A logical model of consciousness on an autonomously adaptive system. *Int. J. Mach. Conscious.* 1, 235–242. doi:10.1142/S1793843009000219

Kinouchi, Y., and Kato, Y. (2013). A model of primitive consciousness based on system-level learning activity in autonomous adaptation. *Int. J. Mach. Conscious.* 5, 47–58. doi:10.1142/S1793843013400040

Kinouchi, Y., and Mackin, K. (2015). An approach for the binding problem based on brain-oriented autonomous adaptation system with object handling functions. *Proc. Comput. Sci.* 71, 76–84. doi:10.1016/j.procs.2015.12.208

Kosslyn, S. (1994). *Image and Brain*. Cambridge: MIT Press.

Kutomi, O., and Hori, M. (2014). The molecular mechanisms of intraciliary supply system and ciliary movements by the studies on *Paramecium cilia*. *Jpn. J. Protozool* 47, 13–27. (In Japanese).

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.* 1, 1–40.

Libet, B. (2004). *Mind Time: The temporal Factor in Consciousness*. Cambridge: Harvard University Press.

Mannella, F., Mirolli, M., and Baldassarre, G. (2016). Goal-directed behavior and instrumental devaluation: a neural system-level computational model. *Front. Behav. Neurosci.* 10:181. doi:10.3389/fnbeh.2016.00181

Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* 301, 229–232. doi:10.1126/science.1084204

Meyer, K., and Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends Neurosci.* 32, 376–382. doi:10.1016/j.tins.2009.04.002

Mischel, W. (2014). *The Marshmallow Test: Understanding Self-control and How to Master It*. London: Bantam Press.

Patterson, D., and Hennessy, J. (1994). *Computer Organization and Design*. San Francisco: Morgan Kaufmann.

Pike, N. (1967). Hume's bundle theory of the self: a limited defense. *Am. Philos. Q.* 4, 159–165.

Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition* 80, 127–158. doi:10.1016/S0010-0277(00)00156-6

Ranzato, M., Huang, F. J., Boureau, Y. L., and LeCun, Y. (2007). "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, CVPR'07. IEEE Conference on*, Minneapolis, 1–8.

Reggia, J. (2013). The rise of machine consciousness: studying consciousness with computational models. *Neural Netw.* 44, 112–131. doi:10.1016/j.neunet.2013.03.011

Reggia, J., Katz, G., and Davis, G. (2018). Humanoid cognitive robots that learn by imitating implications for consciousness studies. *Front. Robot. AI*. 5:1. doi:10.3389/frobt.2018.00001

Rizzolatti, G., Cattaneo, L., Fabbri-Destro, M., and Rozzi, S. (2014). Cortical mechanisms underlying the organization of goal-directed actions and mirror neuron-based action understanding. *Physiol. Rev.* 94, 655–706. doi:10.1152/physrev.00009.2013

Scimeca, M., Steven, L., and Franconeri, S. (2015). Selecting and tracking multiple objects. *WIREs Cogn Sci* 6, 109–118. doi:10.1002/wcs.1328

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 5, 9. doi:10.1038/nature16961

Sligte, I., Steven Scholte, H., and Lamme, V. (2009). V4 activity predicts the strength of visual short-term memory representations. *J. Neurosci.* 29, 7432–7438. doi:10.1523/JNEUROSCI.0784-09.2009

Sligte, I., Vandenbroucke, A., Steven Scholte, H., and Lamme, V. (2010). Detailed sensory memory, sloppy working memory. *Front. Psychol.* 1:175. doi:10.3389/fpsyg.2010.00175

Smith, J. (2017). *Self-Consciousness*. The Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/fall2017/entries/self-consciousness/ (Accessed: November 13, 2017).

Spruston, N. (2008). Pyramidal neurons: dendritic structure and synaptic integration. *Nat. Rev. Neurosci.* 9, 206–221. doi:10.1038/nrn2286

Stalling, W. (2005). *Operating systems, Internals and Design Principles*. Upper Saddle River: Pearson Prentice Hall.

Stuart, G., and Spruston, N. (2015). Dendritic integration 60 years of progress. *Nat. Neurosci.* 18, 1713–1721. doi:10.1038/nn.4157

Tani, J. (2017). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. New York: Oxford University Press.

Tian, L., Thalmann, N., Thalmann, D., and Zheng, J. (2017). The making of a 3D-printed, cable-driven, single-model, lightweight humanoid robotic hand. *Front. Robot. AI*. 4:65. doi:10.3389/frobt.2017.00065

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Arch. Ital. Biol.* 150, 290–326.

van der Velde, F. (2013). Consciousness as a process of queries and answers in architectures based on in situ representations. *Int. J. Mach. Conscious.* 5, 27–45. doi:10.1142/S1793843013400039

von Neumann, J. (1945). *First Draft of a Report on the EDVAC*. Available at: https://library.si.edu/digital-library/book/firstdraftofrepo00vonn (Accessed: November 13, 2017).

Xu, Y., and Chun, M. (2009). Selecting and perceiving multiple visual objects. *Trends Cogn. Sci.* 13, 167–174. doi:10.1016/j.tics.2009.01.008

Zorpette, G. (2017). *The Benefits of Building an Artificial Brain*. IEEE Spectrum. Available at: https://spectrum.ieee.org/computing/hardware/can-we-copy-the-brain (Accessed: May 31, 2017).

Conflict of Interest Statement: 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.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 04 April 2018
doi: 10.3389/frobt.2018.00029

From Focused Thought to Reveries: A Memory System for a Conscious Robot

Christian Balkenius[1], Trond A. Tjøstheim[1], Birger Johansson[1] and Peter Gärdenfors[1,2]*

[1] Lund University Cognitive Science, Department of Philosophy, Lund University, Lund, Sweden, [2] University of Technology Sydney, Ultimo, NSW, Australia

We introduce a memory model for robots that can account for many aspects of an inner world, ranging from object permanence, episodic memory, and planning to imagination and reveries. It is modeled after neurophysiological data and includes parts of the cerebral cortex together with models of arousal systems that are relevant for consciousness. The three central components are an identification network, a localization network, and a working memory network. Attention serves as the interface between the inner and the external world. It directs the flow of information from sensory organs to memory, as well as controlling top-down influences on perception. It also compares external sensations to internal top-down expectations. The model is tested in a number of computer simulations that illustrate how it can operate as a component in various cognitive tasks including perception, the A-not-B test, delayed matching to sample, episodic recall, and vicarious trial and error.

Keywords: working memory, semantic memory, computational model, episodic memory, consciousness

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Anders Lennart Green,
Södertörn University, Sweden
James A. Reggia,
University of Maryland, College Park, United States

****Correspondence:***
Christian Balkenius
christian.balkenius@lucs.lu.se

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *31 October 2017*
Accepted: *07 March 2018*
Published: *04 April 2018*

Citation:
Balkenius C, Tjøstheim TA, Johansson B and Gärdenfors P (2018) From Focused Thought to Reveries: A Memory System for a Conscious Robot. Front. Robot. AI 5:29. doi: 10.3389/frobt.2018.00029

1. INTRODUCTION

1.1. The Inner World

Consciousness is not unitary but involves several kinds of components. The most fundamental component may be the emotional tone of the current state of the mind (Damasio and Marg, 1995). However, in this article, we will not consider emotions but focus on sensations that are the immediate sensory impressions, perceptions that are interpreted sensory impressions, and imaginations (or images) that are not directly governed by sensory impressions (Humphrey, 1992; Gärdenfors, 2003). After emotions, this is presumably the evolutionary order in which the different functions appear. Even for simple organisms, the sensory organs generate sensations. Perceptions require more advanced cognitive processing. The main function of perceptions is to provide information about the animal's environment. Imaginations also require that sensations can be suppressed. The planning behavior of mammals and birds suggests that they have imaginations that concern entities not currently present in the environment.

On the first level, consciousness contains sensations. Our subjective world of experiences is full of them: tastes, smells, colors, itches, pains, sensations of cold, sounds, and so on. This is what philosophers of mind call qualia.

On the second level, an organism that in addition to bodily sensations is capable of representing what is happening at a distance in space or in time will be better prepared to act and thus improve its chances of survival. Several processes in the brain add new information to what is given by the sensations. This holds especially for the visual modality. For example, an object is perceived to have contours, but in the light that is received by the retina, there is nothing corresponding to such structures—this information is constructed by the visual process. By filling in extra information, perceptions help us choose more accurate actions.

On the third level, that of imaginations, sensory input is not used to trigger the filling-in processes, but they are initiated by inner mechanisms. An organism with imaginations can generate a prediction of the consequences of a particular action. Such simulations constitute the core of planning processes. The mechanisms involved in performing an action are the same as those in imagining a performance.

Imagining an action presupposes that the current sensations can be blocked, lest they conflict with the imagination. Glenberg (1997) writes that imaginations put reality in quarantine. The blocking is part of the executive functions mediated by the frontal lobes of the cortex. Glenberg (1997) distinguishes between "automatic" and "effortful" memory. The automatic memory is used to turn sensations into perceptions. For example, finding your way at home in the dark involves blending your limited sensations with your memories.

The effortful memory is used to create imaginations. What is called remembering is a special kind of image that is judged to correspond to an actual event. Effortful memory is also necessary for fantasies: a sphinx cannot be imagined unless you have previous memories of lions and humans.

Perceptions and imaginations taken together generate the "inner world" of an organism. Such an inner world is valuable from an evolutionary perspective. Craik (1967) writes: "If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which are the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react on a much fuller, safer and more competent manner to the emergencies which face it." For an organism with an inner world, actions are generated from a represented goal, rather than directly from the sensations (Jeannerod, 1994). This means that an organism that has imaginations has large advantages to one who must solve a problem by trial and error that can both be very inefficient and lead to dangerous situations. The inner world makes is possible for the organism to simulate different actions and evaluate their effects. Such simulations allow it to select the most appropriate action. Early evidence for such a process was presented by Tolman (1948), who showed that the searching behavior of rats in mazes is best explained by assuming that they have a "spatial map" as part of their imaginations.

An inner world is a *sine qua non* for consciousness. In this article, we will use two memory tests from research on infants as minimal criteria for deciding whether a system has an inner world: (1) exhibiting object permanence and (2) passing the "A-not-B" test (Piaget, 1954).

A child who exhibits object permanence understands that objects continue to exist even when they are not directly perceived. Piaget (1954) studied this by observing infants' reaction to when a favorite object was hidden, say, under a pillow. According to him, object permanence develops between 4 and 8 months of age, but some researchers claim that it may develop earlier (Bower, 1974). Without object permanence an infant would not be able to identify an object or a person over time. It is considered to be a method for evaluating working memory in young infants.

In an A-not-B test, a toy is hidden under box A that is within the reach of an infant. The infant searches for the toy under box A and finds the toy. The hiding is then repeated several times. Then, in the test, the toy is hidden under box B that also is within the infant's reach. Infants between 7 and 10 months typically make a perseveration error, looking under box A even though they saw the toy being hidden under box B. This behavior indicates that the infants have limited object permanence. When infants are 12 months or older, they normally do not make this error.

In this article, we present a novel memory system that supports the minimum operations for a conscious robot with the properties described earlier. The main function of this memory system is to move some cognitive operations into an inner world, and more importantly, to allow the inner world of the cognitive system to coevolve with the external world in such a way that it can generate expectations as those involved in object permanence and the A-no-B test. These expectations can be used in decision-making, to detect changes in the external world, and to direct attention. Furthermore, by allowing the inner world to become decoupled from external input, it can produce chains of "thoughts" based on semantic and episodic relations. Such chains can range from replay of previous episodes to novel combinations of previous experiences. In machines, an inner world in general and object permanence in particular promises to enable more robust goal directed action, visual search, and even planning.

We take a developmental robotics approach (Asada et al., 2009), and first want to model memory processes of the young infant, and later approach more complex abilities. Our goal here is to show how the proposed memory model supports many cognitive functions that are central to a conscious intelligent robot and to suggest that the model could form an important component of a larger cognitive architecture that will be tested in a robot in the future.

1.2. Models of Memory

One of the most canonical models of associative memory is the Hopfield network (Hopfield, 1982, 1984). The Hopfield network consists of a set of nodes connected by associations of varying strengths that store a set of patterns. The network operates as a content addressable memory where an incomplete activation pattern over the nodes will recall a complete stored pattern. An interesting aspect of the network is that it is possible to define an energy function that described every state of the network. It can be shown that the network changes its state in such a way that it decreases the energy of the whole system until it ends up in a local energy minimum. The minima of the energy function correspond to the stored memories. These states are attractors for the system in the sense that any initial state will move toward one these states. These types of networks lend themselves to model both perception and semantic memory but can also be extended to handle episodic associations by introducing delays on associations (Sompolinsky and Kanter, 1986). These properties are central to the model that we develop below and are used to process both semantic and episodic memories and to form associations that binds stimuli to places.

Cognitive operations also require working memory mechanisms and many computational models have been proposed. They emphasize different aspects of the working memory system, such as spatial map formation (Blum and Abbott, 1996), serial order recall (Page and Norris, 1998; Burgess and Hitch, 1999; Botvinick and Plaut, 2006), perseveration and distractibility (Kaplan et al., 2006), gating, action selection, and reinforcement learning (Ponzi, 2008), or sequence generation (Verduzco-Flores et al., 2012). One early computational model of working memory was proposed by O'Reilly et al. (1999). This model includes a prefrontal system that maintains contextual information that is used to bias different processes in the rest of the model. This is combined with a fast learning model of the hippocampus. Similar models were also described by Cohen et al. (1990) and Miller and Cohen (2001).

Focusing on the control aspect of working memory, Sylvester et al. (2013) describe a working memory system that controls the flow of information by opening and closing a network of gates. This system was used to do working memory cycling and comparison and was structured to adequately respond to *n*-back type tasks. Building on the gate paradigm, Sylvester and Reggia (2016) showed how a visual input could be associated with a location in the visual field to perform a card matching task. Both these systems rely on an instruction sequence memory (ISM) that can be programmed with sequences of gate configuration so as to respond adequately to the task at hand. The ISM consists of a Hopfield network (Hopfield, 1982) that can store attractor sequences by a mechanism of Hebbian learning (Hebb, 1949).

Moving away from cognitive and brain inspired models, more abstract neural network models have also begun to incorporate association mechanisms. For example, there has been a growing interest in adding external memory systems to deep-learning networks. In conventional deep-learning models, the memory of the network is stored implicitly in the entire network, in the form of unit weights. Hence, it is hard to store particular associations in such structures. This has prompted research into architectures that add external memory modules, allowing activation patterns to be stored alongside other data, such as labels, words, or sounds.

Most such memory modules, like the neural Turing machine (Graves et al., 2014) and the differentiable neural computer (Graves et al., 2016), evolvable neural Turing machine (Lüders et al., 2017; Parisotto and Salakhutdinov, 2017), have a form of key—value mechanism where the key is typically the output from another network structure like a convolutional or recurrent net. Depending on the sophistication, such memory modules can update based on evidence, learn ordering patterns, or supply answers to queries (Weston et al., 2014; Chen et al., 2015).

The memory system we propose here shares some properties with these models but is different in that it explicitly aims at roughly reproducing the properties of specific brain regions.

2. THE MEMORY SYSTEM

This section describes the main components of the memory system and their functions. The model includes three interacting neural networks that roughly correspond to the ventral, dorsal, and prefrontal areas of the cortex (**Figure 1**). First, an identification network transforms sensations into perceptions; second, a localization network codes the spatial location of an object; and, third, a working memory network retains recently activated patterns over time.

2.1. Identification Network

The first component is the identification network that learns different stimuli as collection of stimulus properties. It corresponds to the WHAT system of the ventral cortex as proposed by Mishkin et al. (1983) and Goodale and Milner (1992). The part of this system that is included here can be sees as the highest level in a sensory processing hierarchy generating perceptions. It operates as a content addressable memory and recalls complete patterns based on partial inputs. We also assume that it generates top-down influence on sensory processing and interacts with value systems (Balkenius et al., 2009), but we do not model that here.

The identification, or WHAT, system is implemented as a fully connected network (see Appendix in Supplementary Material). This allows the network to settle into attractors that represent different memory states. In addition to the usual dynamics, we also include a mode of synaptic depression (Abbott et al., 1997; Tsodyks et al., 1998). This leads to a latching dynamics where the network can autonomously transition between different attractors (Lerner et al., 2010, 2012, 2014; Aguilar et al., 2017). This

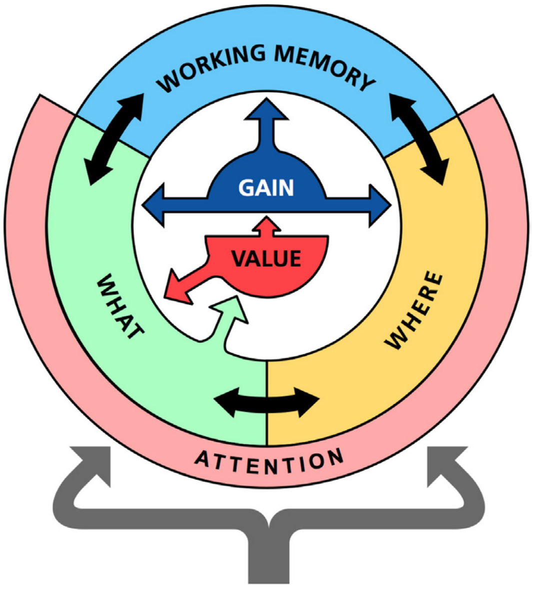


FIGURE 1 | Overview of the memory model. The memory model consist of three main parts: the identification network (WHAT), the localization network (WHERE), and a prefrontal working memory network (WORKING MEMORY). Each network is modeled as a recurrent neuronal network with similar design but with slightly different dynamics. In addition to internal recurrent connections, there are also temporal associations that can read out sequences of states in memory. The identification and localization networks also include an attention component that detects novel external stimuli and compares expected to actual inputs to potentially generate surprise signals. The identification network communicates with value system (VALUE). All processing is under the influence of a gain modulation system (GAIN) that controls the randomness of the state transitions in memory.

can be seen as free associations between the stored memory states (Russo et al., 2008; Akrami et al., 2012; Russo and Treves, 2012).

Furthermore, the identification network includes a comparator that compares the sensory input to the corresponding attractor state (Balkenius and Morén, 2000). Any stimulus or attractor component that differs contributes both to a total measure of surprise and to a feature-specific surprise that includes the parts of the sensory input that does not match the attractor state.

The current memory state is assumed to tune the attention system toward stimuli that match the state. For example, a state coding for the color red would tune the attention system to look for red objects in a way akin to the feature integration theory of attention (Treisman and Gelade, 1980). The identification network is thus assumed both to influence attention through top-down expectations, and to be influenced by bottom-up perceptual processes.

2.2. Localization Network

The second component is the localization network, or WHERE system. It parallels the functions of the parietal cortex (Andersen et al., 1985) and the hippocampus (Smith and Milner, 1981). Its role is to maintain a specific code for each possible location in the environment. This code is assumed to be activated when we look at a particular location.

It is similar to the identification network except that its activity is constrained by a winner-take-all-rule that implements the constraint that only one place is actively represented at each time. Associations between the identification and localization components allow the memory system to store bindings between places and objects. By associating each perceived object with its own individual location, the memory system avoids the binding problem where properties of different stimuli are mixed up in the network (ref). Another role of the localization network is that it increases the storage capacity of the identification component and avoids spurious attractors. The reason for this is that the localization codes are orthogonal for each location.

Like the identification network, this part of the memory model participates in both bottom-up and top-down processing. When we attend a particular location, the code for that location is activated in the localization network. Similarly, when a location code is activated by internal processes, it will influence attention and make us more likely to look at the coded location.

2.3. Working Memory Network

The final component is a "prefrontal" working memory (Fuster, 2009). The function of this network is to allow memories "stored" in working memory to be more easily recalled than other memories. According to our model, the actual working memories are not stored in the prefrontal system. Instead, the working memory function is the result of the interaction between prefrontal and sensory cortical areas. The working memory activation thus does not contain any sensory attributes although it is able to recall such attributes in the identification and localization networks (Lara and Wallis, 2015).

To allow the limited working memory to store any possible object–place binding, the nodes of this network are recruited when needed. The process is similar to that of an ART network (Grossberg, 1987), but less elaborate. The recruited nodes maintain an active state as long as the working memory is active. It is well known that prefrontal working memory cells operate in this way and allows for persistent activation during a memory period (Wang, 2001; Curtis and D'Esposito, 2003).

Each active working memory node can potentially influence the states of the identification and localization networks. Which node is allowed to do this depends on both the similarity of its learned input pattern and the current state of the complete system as well as the activity level of the node itself. The result of this mechanism is that a partial cue will recall the most recent state that is similar to the input.

The influence from the working memory network on the rest of the system involves both excitation and inhibition and can be likened to the inhibitory control exhibited by the prefrontal cortex (Fuster, 2009). Once a working memory node has been selected, it will promote the coding of its stored memory and inhibit other stimulus components (Desimone and Duncan, 1995). This can be seen as a top-down modulation of the states in the identification and localization networks (Gazzaley and Nobre, 2012). It can also indirectly control spatial attention through the localization network (Corbetta and Shulman, 2002).

2.4. Predictive Associations

In addition to the associations between the three networks, the memory system also contains predictive associations that work over time to predict the next state based on the current one. When allowed to run freely, these temporal associations will make the complete system transition between stable attractors over time in a way akin to daydreaming. When there is no input to the memory system, it will instead recall and internally play previously experienced sequences. As we will show below, this mechanism can be put to good use in choosing between different actions depending on their expected outcome. The predictive associations are learned in the same way as other associations except that there needs to be a delay between the activation of the two nodes that will be associated together. This will make the network to learn an association to the current state from a previous state of the network. The delay during learning is mirrored in a delay in the association that will be used to read out the prediction in the future.

2.5. Modes of Operation and Metaparameters

There are several parameters that can influence the operation of the memory system. The first is the level of noise. Memory transitions are highly dependent on the noise level and with sufficient noise; the state of the memory system will jump randomly between the different attractors. A moderate amount of noise allows the memory state to take new directions without being completely random, and a lower level makes the memory system more likely to stay in the same state for a longer time or to follow precise episodic memories.

In the brain, the locus coeruleus is believed to adjust the sensitivity to noise. This is a general arousal system and the main source of noradrenergic input to most of the brain. It has been suggested that the locus coeruleus, instead of changing the noise level, changes the response to noise by modulating the gain of cells involved in decision processes (Chance et al., 2002; Aston-Jones and Cohen, 2005; Donner and Nieuwenhuis, 2013; Eldar et al., 2013). Doya (2002) proposed that this should be seen as a metaparameter that allows the randomness of the processing to be controlled.

The second main parameter is the relative influence of the external input and internal expectations in controlling the memory state. The system can run in either in bottom-up mode where the internal state is controlled by external stimuli or in top-down mode where the sequence of memory states is internally produced. It is also possible to combine bottom-up and top-down processing. This allows the internal expectations to be compared with external stimuli and to make the system surprised when expectations are not met. Such a comparison also has an additional role. When there is a sufficiently large mismatch between the sensory input and the internal state, the memory system will be reset to allow the novel stimulus to quickly be coded in the different memory networks.

In the following sections, we apply the general memory system to a number of tasks and show how it can form the basis for many fundamental cognitive tasks. In these simulations below, the metaparameters were set heuristically to allow the model to show the desired properties in each case. When the memory system is used as a part in a complete architecture, these parameters are assumed to be learned for each particular task.

3. FROM SENSATION TO PERCEPTION

The role of perception can be seen when considering the well-known Kanizsa triangle (Kanizsa, 1976) (**Figure 2**). Our perceptions tell us that a white triangle lies on top of three black circles. Yet in the figure, there are no lines marking off the sides of the triangle from the white surroundings. The lines are a construction of our brains. There is a mechanism that simulates the existence of lines completing the segments of the circles.

Examples like this show that we have plenty of processes that complement the signals provided by the senses. Such complementations create the representations with which memory works—the perceptions, since what we remember is not only that which is presented by our sensory receptors but also that which is recreated, i.e., represented, by the filling-in processes. Here, we only consider a network with identical nodes and connections, but the reasoning is equally valid for more complex network. For example, Månsson (2006) developed a complex network that fills in contours in the Kanizsa triangle using a range of neuron models with different properties.

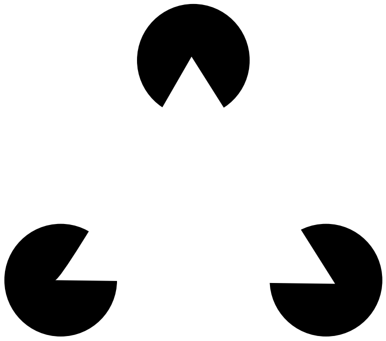


FIGURE 2 | The Kanizsa triangle.

In **Figure 3**, we illustrate how the pattern completion mechanism operates in the memory system. The system has learned three patterns, one of which is the letter L. When parts of the L are activated, the identification network will fill in the missing parts of it. In the figure, there are three stored patterns represented by different colors.

4. OBJECT PERMANENCE

A cat chasing a mouse that runs in behind a curtain can predict that it will come out the other side. So the cat can draw conclusions about the mouse even when it is receiving no direct signals from its senses. Such behavior presumes the cognitive ability called object permanence by Piaget (1954). This implies that the cat retains some kind of representation of the mouse even when its sensory impressions of the mouse are gone. The cat has expectations concerning the mouse.

Various studies of animals show that all mammals, birds, and octopuses possess object permanence. These organisms thus enjoy one more way to build in knowledge about the future in their consciousness. Object permanence is not innate, but it must be learned.

To test the memory model for its capacity to handle object permanence, we simulated two types of memory tasks. In both cases, the system is first presented with three objects X, Y, and Z.

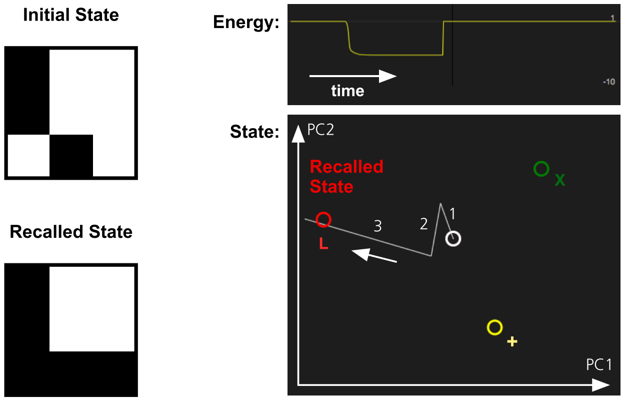


FIGURE 3 | Pattern completion in the memory system. The memory has learned three patterns, L (red), X (green), and + (yellow). The partial activation of the L-pattern will make the memory system recall the complete pattern. The graph at the top right shows how the energy of the memory state decreases as the pattern in recalled. The graph below shows the memory state projected on a two-dimensional space defined by the first two principal components (PC1 and PC2) of the stored memory patterns. The graph shows the transition between an initial inactive state (white) and the recalled state (red). The numbers and arrow indicate the sequence of the different transitions.

Each at its own location A, B, and C. In the first simulation, we tested if the memory system could recall the location of objects that it had previously seen (**Figure 4**). The memory was first cued with object X. This makes the memory state transition to the attractor for X. At the same time, the localization part of the memory system activates the location A that is associated with X. The locations are recalled for object Y and Z as well. Finally, we tested what is the result if we cue the memory with a stimulus that is similar to both X and Y. Here, we used an input pattern that contained only components that were shared by both objects. As can be seen in **Figure 4**, the memory state transitions to the attractor for object Y. The reason for this is that Y is more strongly coded in the working memory since it was seen more recently than X. In addition to showing the role of the working memory, this is also an example of pattern completion. The initial pattern is similar to both X and Y, and the memory state first moves toward a place between X and Y, before turning toward Y as more properties of Y are filled in.

In the second simulation, we tested whether the memory system can recall objects by being cued with locations. The results of this simulation are shown in **Figure 5**. When a location is cued, the memory state transitions to the attractor for the corresponding object illustrating that the memory system has formed expectations of which object is where.

The simulations show that the memory system can learn what object to expect at a particular location. Together with the comparator that compares expected and actual input, this allows the system to become surprised if expectations are not met (cf. Balkenius and Morén, 2000). It can also recall where it has seen an object. Such information can be used to determine where to search for an object and to direct the gaze while looking for it. The memory system thus has the essential properties needed for object permanence.

5. A-NOT-B

Another way to address object permanence is to run the A-not-B experiment on the memory model. To test if the memory system would make the A-nor-B error, we simulated the A-not-B task under two conditions. In the first, the output gain of the working memory system was low to simulate a brain at an earlier stage of development. In the second, the working memory gain was set at full strength. The system was first trained by repeatedly showing object X at location A. In the second step, we simulated moving object X to location B. This results in two stored memories in long-term memory, a stronger one that associates X with location A and a weaker one that associates X with location B.

To test the system, we activate the pattern for X in the WHAT system and allow the system to activate a location code in memory. When the working memory is turned off, the stronger association will win, and the system will recall location A (**Figure 6**). However, when the working memory system is turned on, the result is different. In this case, the working memory will

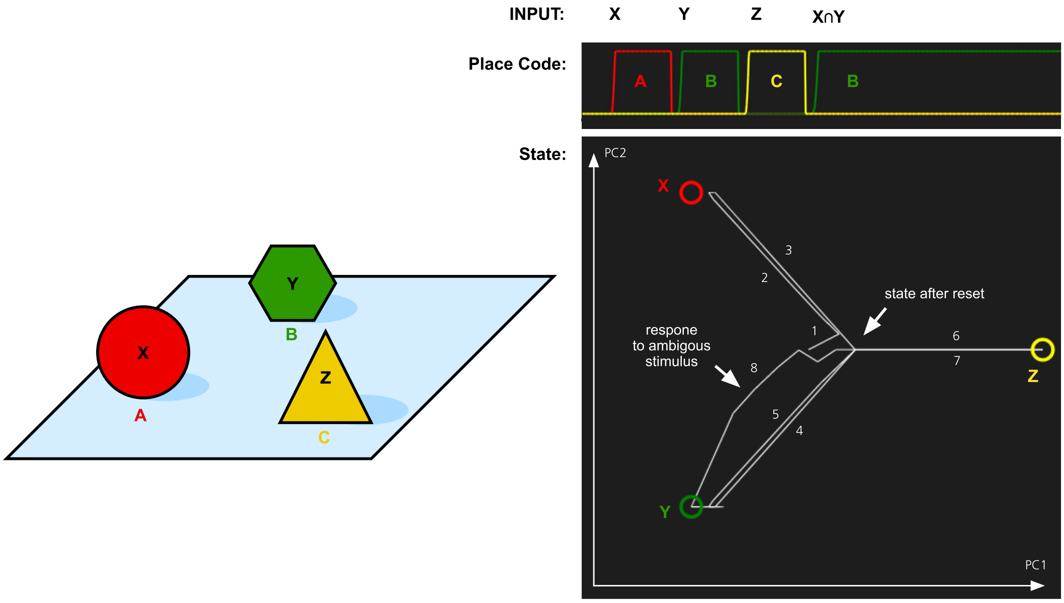


FIGURE 4 | Simulation of place recall. Left: A scene with three objects X, Y, and Z at three places A, B, and C. The memory system was initially trained on these three objects. Right: The graph at the top shows the activation of the localization network when each object is used as input. Finally, an input pattern that consists of the overlapping parts of X and Y is used as input. This stimulus is equally similar to X and Y and thus ambiguous. The result is that the most recently attended place with an object similar to the input is recalled, that is, B. At the same time, the activity pattern in the identification system restores the complete pattern for Y. The graph at the bottom left shows transitions through the memory space. The image shows the memory state over time plotted in a two-dimensional space generated by the first two principal components (PC1 and PC2) of the attractor states. The circles represent the memories of X, Y, and Z, and the line shows how the memory state transitions between the memories as a response to the different input and the numbers show the order of the different transitions. The center of the image where all lines meet corresponds to the empty memory state after reset.

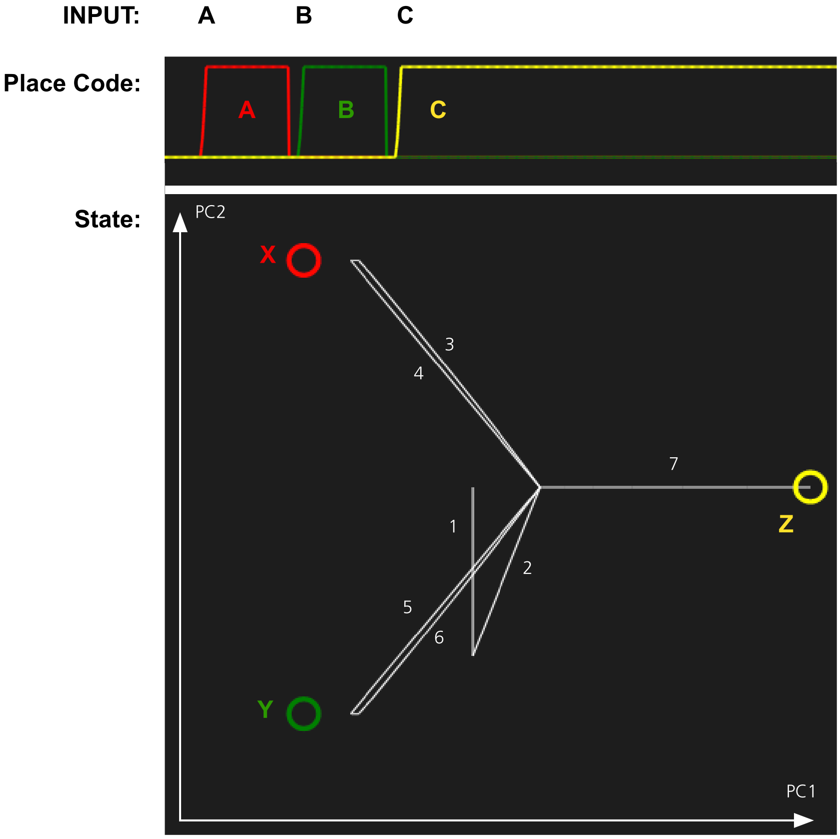


FIGURE 5 | Simulation of the object recall for the scene in **Figure 4**. The memory system was first shown three combinations of objects and place: AX, BY, and CZ. Next it is cued with each of the locations A, B, and C. The graph at the top shows the activation of each place code over time. The graph at the bottom shows the path through the memory space as each location is cued. The state is initially wandering, which results in transitions 1 and 2 just before the system is cued with A.

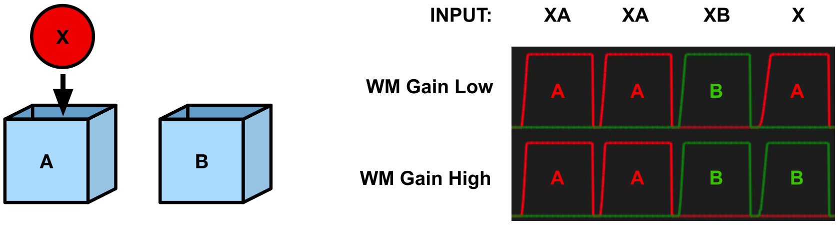


FIGURE 6 | Simulation of the A-not-B task. Left—The object X and the two boxes A and B. Right—The graphs show the activation of the place code with low or high working memory gain for place A (red) and B (green), respectively, as response to different inputs. X represents the object stimulus, and A and B represent the two boxes. The final input X corresponds to the questions "Where is X?" With an undeveloped prefrontal cortex (low working memory gain) the model replies A. With a developed prefrontal cortex (high working memory gain), the model replies B.

remember each perceived stimulus. Every time a new stimulus is perceived, a new node in working memory will be activated while the activity in the remaining working memory nodes will decay slightly. As a consequence, a number of stimuli can be held in working memory at the same time. When a pattern is activated in the WHAT or WHERE components, the working memory cooperates to fill in missing information. Here, the perception of the stimulus X will recall the most recent activation containing X and read out its location B, thus avoiding the A-not-B error (**Figure 6**).

The performance of the models can be related to the serial position effect (Murdock, 1962). The initial error can be seen as a primacy effect, where the initial location of the object is stronger in memory as a result of multiple presentations. The avoidance of the error can be seen as the results of a recency effect, where the most recent location is more easily recalled. This view is in line with the model by Munakata (1998) that suggests that the A-not-B error is a result of competition between latent and active memory traces. However, the behavior of the model is different from the usual recency effect since it depends on a working memory component and not on short-term memory.

Our results fit well with findings suggesting that working memory is a driving force in cognitive development (Kail, 2007). An alternative theory of the A-not-B error is that it depends on the strength of the initially reinforced response to search at A (Diamond, 1998). We do not exclude that such a factor could also be involved, but our simulation shows that a working memory explanation may be sufficient. However, the working memory here influences the rest of the system by inhibiting the incorrect location, and similar mechanisms could presumably be used to inhibit an incorrect response in a similar way to an incorrect location.

6. DELAYED MATCHING TO SAMPLE

The delayed matching to sample task (DMTS) is a variant of more general delayed response tasks (Rodriguez and Paule, 2009). Such tasks involve the presentation of stimuli, followed by a delay where no stimuli are given. The original stimulus is then presented along with one or several choice options, and the subject is required to choose which matches the original.

The task can be varied in difficulty by changing the delay time, or by altering the number of options to choose among during the response. Distractors may also be introduced to affect subjects' ability to maintain attention and to impair working memory capacity (Rodriguez and Paule, 2009). Lesion studies in monkeys (Gaffan and Weiskrantz, 1980) indicate that the prefrontal and inferior temporal cortices are involved in DMTS tasks. Specifically, performance for tasks with visual stimuli is impaired after a higher visual area of the inferior temporal cortex has been damaged. Lesioning the prefrontal cortex appears to reduce the delay after which a correct response can be made but does not impair successful completion as such (Mishkin and Manning, 1978).

The configuration of the visual stimuli may take different forms, depending on which specific aspect of memory is under scrutiny. Sawaguchi and Yamane (1999) used a white square presented at one of four peripheral positions, placed equidistantly about a central focus point to study spatial memory. Tanji and Hoshi (2001) used a more complex setup with three cues placed in a pyramid pattern, each showing either a circular or triangular shape. This was used to study behavioral planning based on shape or location matching. Other variations of the DMTS task have been used to study color matching (Mikami and Kubota, 1980; Giurfa et al., 2001), movement matching (Ferrera et al., 1994), and horizontal vs. vertical orientation matching (Giurfa et al., 2001). The simplicity of the task makes it suitable for studying memory effects across various species, including humans (see, e.g., Daniel et al. (2016) for a review).

Using our memory model, we simulated a delayed matching-to-sample task (**Figure 7**). The system is first presented with a sample stimulus X that it will store in working memory. After a delay period, a comparison stimulus, X or Y, is presented. For each stimulus, the working memory network will read out the remembered stimulus and compute the match to each of the comparison stimuli. We assume that there exists a mechanism external to the memory system that selects the stimulus that generates the least surprise.

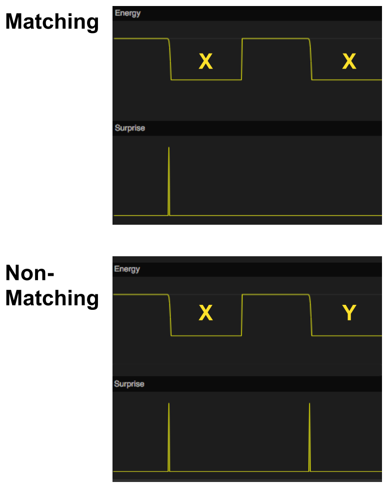


FIGURE 7 | Simulation of delayed matching to sample (DMTS). In the top graph, the stimulus X is first shown as a sample stimulus and is subsequently followed by X again as comparison stimulus. There is no surprise signal the second time X is shown, indicating that the model recalls that it has seen this stimulus before. In the bottom graph, the sample stimulus X is followed by comparison stimulus Y instead. In this case, there is a surprise signal for the non-matching stimulus. The energy function is used to show the timing of the stimuli.

Our simulation shows that the memory system has the necessary memory functions for a delayed matching-to-sample response.

7. DAYDREAMING AND EPISODIC RECALL

Two possible mechanisms are involved in producing transitions between attractors. The first is the noise in the system that can kick the network out of an attractor if it is strong enough. The second mechanism is synaptic depression that weakens synapses that are involved in maintaining the current attractor. This has the effect of eventually making the state wander away from the attractor. A possible interpretation is that this is what occurs when the attentional system is not engaged, which makes the memory system enter a state of daydreaming where it can wander freely. The mind wandering produced by the model does not have any function but is instead a natural consequence of the function of the memory system. This is in line with the view presented by Mason et al. (2007) who suggest that the mind wanders "simply because it can."

Herrmann et al. (1993) distinguish between semantic and episodic transitions in neural networks. Semantic transitions occur between states that are semantically related and are caused by synaptic depression that moves the state away from one attractor in favor of another one with overlapping activation pattern. Episodic transitions, on the other hand, are caused by predictive temporal associations (Sompolinsky and Kanter, 1986).

Figure 8 shows a simulation of semantic associations in the memory system. The system was first trained with three patterns X, Y, and Z where X and Y share some features, Y and Z share some other features, but X and Z do not share any features. With low noise, the system transitions randomly between X and Y, and between Y and Z, but not between X and Z. With a higher level of noise, the transitions occur between all states. Finally, with no noise, the system returns to the same state after synaptic depression. Although we want to like this wandering to daydreaming, it is obviously limited to combinations of states that the network has previously experienced.

Figure 9 shows a simulation of episodic recall in the memory system. The system was first trained with two sequences of stimuli: X, Y, Z and P, Q, R. When presented with X as an input, the memory system will read out the sequence X, Y, Z (**Figure 9A**). Similarly, for an input P, the sequence P, Q, R will be produced. When the noise level is increased, the episodic recall will sometimes transition from Z to Q, producing a novel sequence X, Y, Z, Q, R (**Figure 9B**). This shows how the memory system can combine two episodes into a novel imagined episode. The evolutionary value of such reveries is that they allow the memory system to generate new combinations of memories that can form the kernels for new plans. Some of these plans can be tried out at later occasions. Hence, the same mechanism that produces daydreaming can be seen as an element in a generate-and-test procedure.

8. VICARIOUS TRIAL AND ERROR

If an agent has an internal model of the world, it can make simulations of the consequences of actions (Craik, 1967). Redish (2016) proposes that animals internally simulates the outcomes of different choices before making the choice in the external world. As noticed by Muenzinger (1938) and Tolman (1939), rats look back and forth at different alternatives at a choice point. A rat that has to choose whether to go left or right in a maze can use its episodic memory to simulate selecting the left or the right path (**Figure 10**). The episodic memory recall described earlier is ideally suited for this process. By cueing the memory system with the stimulus A to the right, the sequence of moving through A, B, and C will be simulated internally. When looking right to see X, the sequence X, Y, Z, G will be produced instead. Since this sequence leads to the goal, the rat can now chose to go right.

Figure 10 shows a simulation of vicarious trail-and-error in a simple maze. The memory system has first experienced moving through the maze along two different routes. The first consists of locations A, B, and C which is a dead end, and the other consists of the sequence X, Y, Z, which finally leads to the goal G. At a choice point in a maze, the robot can look left or right, and the memory system is used to imagine the result of select one of the two possible paths. Looking at A, which will read out the sequence A, B, C that does not lead to a goal, and looking at the second alternative X, will read out the sequence X, Y, Z, G, which ends with the goal. This mechanism could be used by a decision

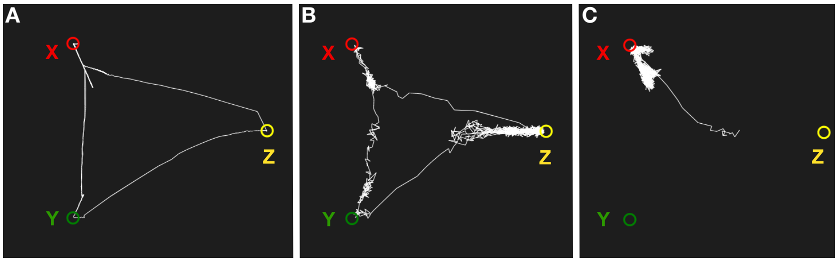


FIGURE 8 | Simulation of mind wandering using semantic associations. **(A)** With low noise, the system will transition between semantically related states as a result of synaptic depression. **(B)** With a higher noise level, the memory system will transition less regularly and can potentially end up in semantically unrelated states. **(C)** With low synaptic depression, the system will move away from an attractor but return back again most of the time.

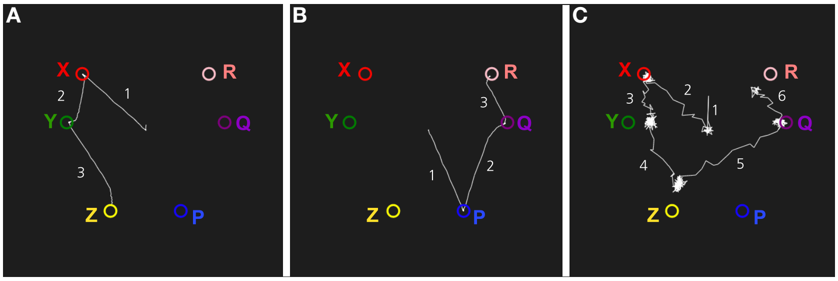


FIGURE 9 | Simulation of episodic recall. The three graphs show transitions between the attractors of the network. **(A)** Recall of the episode X, Y, Z cued by an input X. **(B)** Recall of the episode P, Q, R cued by P. **(C)** A higher noise level produces a novel imagined episode that is a combination of two experienced episodes: X, Y, Z, Q, R.

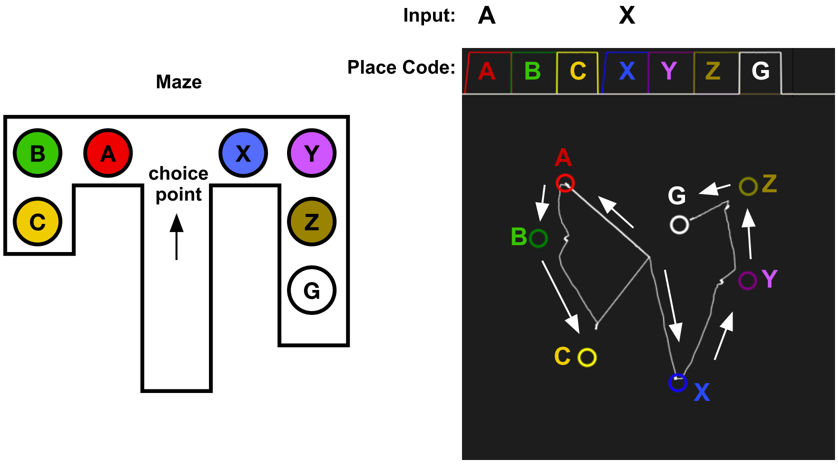


FIGURE 10 | Vicarious trial and error. The memory system is assumed to have learned the sequence of places that are experienced while traveling through the maze. At the choice point, the memory system is used replay the result of choosing A or X. When looking left toward A, the memory cued with A and will start to replay A, B, C. When looking right, the memory system is cued with X which will replay the sequence X, Y, Z, G, which leads to the goal. The graph in the top right shows the activation of the place codes on the localization network. Note that the activation of A and X is slower as they are cued by an external stimulus. The graph in the bottom right shows the transitions through the identification network. The state starts at the center as A or X is received and moves through the states for the different places in the maze. The arrows show the direction of the memory transitions.

mechanism that chooses between alternative actions based on their expected consequences.

Redish (2016) suggests that this type of mechanism is responsible, not only for spatial navigation but also for deliberative processes in general and that the internal schema used to simulate the world is what (Tolman, 1948) would call a cognitive map. This view of the cognitive map is in line with Tolman's original view where the cognitive map did not have to be spatial but could be used for any kind of problem solving. Our proposed memory model can thus operate as a cognitive map that supports elementary planning operations.

9. DISCUSSION

We have introduced a memory model for robots that can account for many aspects of the presence of an inner world, ranging from object permanence, episodic memory, and planning to imagination and reveries. It is modeled after neurophysiological data and includes many parts of the cerebral cortex together with a model of the arousal system. It consists of three main components, an identification network, a localization network, and a working memory network. An important aspect of the model is that the mechanisms that fill in sensations to generate perceptions can be detached from sensory input and run in isolation (Gärdenfors, 2003). This allows for planning mechanisms and for daydreaming that can serve as an investigation of a space of possibilities as a preparation for generating plans.

We propose that a robot equipped with this memory system together with mechanisms for more advanced sensory processing and action selection would have the required cognitive equipment to produce a basic form of consciousness—at least to the extent that it can be tested in behavioral experiments. A fundamental aspect of this model is that consciousness in not something that has to be added to the cognitive system. Instead, it is something that occurs naturally once a memory system is able to fill in sensory information and produce memory transitions over time. This will create an inner world that is used both to interpret external input and to support thoughts disconnected from the present situation.

The memory system can operate either in bottom-up mode, where external input directly controls the internal state, or in top-down mode, where previously experienced episodes control the progression of internal states. The internal flow of thoughts is modeled as transitions between memory states. The randomness of these transitions depends on the input from the locus coeruleus. In one extreme, the memory state is stuck in the current attractor, but when the sensitivity to noise increases, the memory state will start to transition to semantically similar states—also supported by synaptic depression. At the same time, episodic associations between states will make the memory replay sequences of states that it has previously experienced. When the randomness increases further, the memory state can make transitions between increasingly unrelated states. The locus coeruleus input thus acts as reins for focusing thought and thus preventing the system from ending up in galloping reveries.

It is an open question how the randomness of the memory processes should be controlled to optimally utilize the memory system for different tasks. Here, we did not include other parts of a complete system that could operate on the memory system. One interesting addition would be to add a reinforcement learning system that could learn to control the level of noise in the memory system to control transitions between different attractors (Lerner and Shriki, 2014). Such a reinforcement learning system could potentially control the various metaparameters to adapt the memory processing to the task at hand (Doya, 2002).

Another addition would be to allow a reinforcement learning system to control the different memory operations, in particular the storage and read out from working memory. In the current model, working memory is not controlled explicitly but stores every memory state as it occurs. From a developmental perspective, this is a reasonable approach before efficient utilization of working memory has been learned and constitutes a substrate for future learning of internal memory operations.

The memory system presented in this work can be contrasted with that described by Sylvester and Reggia (2016). The main difference between their work and ours is first the employment of gates, and second the inclusion of a discrete control module to sequentially set configurations of those gates. There is also a difference in the way the systems learn. Sylvester and Reggia (2016) explicitly program their system by imposing attractor states on a sequence memory part of the control module. By contrast, our system learns sequences of states from observation. Hence, Sylvester and Reggia (2016) can be likened to a system being taught by a teacher, while our system learns by discovery. Both systems utilize Hopfield networks for storing attractor states and employ forms of working memory. In our case, although the working memory does not store visual patterns as such, only associations between high-level sensory representations. The nature of those representations is arbitrary, but we chose to focus on object identity and location for this work. We do, however, acknowledge the utility of gating mechanisms for learning action sequences and plan to incorporate such mechanisms in future models.

Another important next step will be to test the model on a humanoid robot. We will use visual input from cameras that will be analyzed through a bidirectional deep-learning network before reaching the identification network described here. Similarly, the localization network will receive input that uses a population code for locations in three dimensions in several coordinate systems. A robotic implementation already exists with a minimal version of each of these components, but further development of the sensory processing is needed before the experiments simulated here can be tested in a robot in a natural environment.

When the internal processes meet the external input, the memory system is used to compare expectations against the external world to potentially produce surprise and control action selection. We did not include mechanisms for action selection here, but the output from the comparator of the attention system could easily be used for such selections. For example, to learn delayed matching or non-matching to sample, an action selection system would only have to associate the output of the comparator with selecting to refraining from selecting a particular stimulus. Similarly, to choose the correct path through a maze, the mechanism for vicarious trial and error we demonstrated would need to be interfaced with an action selection mechanism that learns to evaluate alternatives and select the one that leads to the goal. Given that the memory system does most of the work, very little remains to be learned by an action selection system.

Attention plays a crucial role as the interface between the inner and the external world. It directs the flow of information from sensory organs to memory and in the other direction it is responsible for the top-down influences on perception. The internal and external world can be seen as two dynamical systems that can be coupled or decoupled in different ways depending on the state of the organism and the task at hand. This allows the proposed model to bridge the gap between cognition as internal processing and situated cognition. We suggest that during evolution, as well as during the development of an organism, one finds a gradual change from acting in the external environment to operating in an internal world.

When the flow of thought through the inner world is cued by the immediate external stimuli, the memory system is used to evaluate the consequences of different available options. When allowed to flow freely, there need not be any relation between the train of thought and the current situation, but by changing the balance between bottom-up and top-down processing, the system can quickly be dragged back to the present situation. On the other hand, when the bottom-up influence is low, the system will start to daydream and replay experienced episodes or producing novel never experienced episodes by combining memories in new ways. The new combinations can then be used as input to the planning mechanisms. The same mechanisms are thus used both for focused goal-directed thought and for daydreaming and reveries.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/frobt.2018.00029/full#supplementary-material.

REFERENCES

Abbott, L. F., Varela, J., Sen, K., and Nelson, S. (1997). Synaptic depression and cortical gain control. *Science* 275, 221–224. doi:10.1126/science.275.5297.221

Aguilar, C., Chossat, P., Krupa, M., and Lavigne, F. (2017). Latching dynamics in neural networks with synaptic depression. *PLoS ONE* 12:e0183710. doi:10.1371/journal.pone.0183710

Akrami, A., Russo, E., and Treves, A. (2012). Lateral thinking, from the hopfield model to cortical dynamics. *Brain Res.* 1434, 4–16. doi:10.1016/j.brainres.2011.07.030

Andersen, R. A., Essick, G. K., and Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science* 230, 456–458. doi:10.1126/science.4048942

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., et al. (2009). Cognitive developmental robotics: a survey. *IEEE Trans. Auton. Ment. Dev.* 1, 12–34. doi:10.1109/TAMD.2009.2021702

Aston-Jones, G., and Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J. Comp. Neurol.* 493, 99–110. doi:10.1002/cne.20723

Balkenius, C., and Morén, J. (2000). "A computational model of context processing," in *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behaviour*, eds J.-A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat, and S. W. Wilson (Cambridge, MA: MIT Press), 256–265.

Balkenius, C., Morén, J., and Winberg, S. (2009). "Interactions between motivation, emotion and attention: from biology to robotics," in *Proceedings of the Ninth*

International Conference on Epigenetic Robotics, Vol. 149, eds L. Cañamero, P.-Y. Oudeyer, and C. Balkenius (Lund: Lund University Cognitive Studies), 25–32.

Blum, K. I., and Abbott, L. (1996). A model of spatial map formation in the hippocampus of the rat. *Neural Comput.* 8, 85–93. doi:10.1162/neco.1996.8.1.85

Botvinick, M. M., and Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychol. Rev.* 113, 201. doi:10.1037/0033-295X.113.2.201

Bower, T. G. (1974). *Development in Infancy*. San Fransisco: WH Freeman.

Burgess, N., and Hitch, G. J. (1999). Memory for serial order: a network model of the phonological loop and its timing. *Psychol. Rev.* 106, 551. doi:10.1037/0033-295X.106.3.551

Chance, F. S., Abbott, L., and Reyes, A. D. (2002). Gain modulation from background synaptic input. *Neuron* 35, 773–782. doi:10.1016/S0896-6273(02)00820-6

Chen, J., He, J., Shen, Y., Xiao, L., He, X., Gao, J., et al. (2015). "End-to-end learning of LDA by mirror-descent back propagation over a deep architecture," in *Advances in Neural Information Processing Systems*, Vol. 28, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (New York: Curran Associates, Inc), 1765–73.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol. Rev.* 97, 332. doi:10.1037/0033-295X.97.3.332

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201. doi:10.1038/nrn755

Craik, K. J. W. (1967). *The Nature of Explanation*, Vol. 445. Cambridge: Cambridge University Press.

Curtis, C. E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* 7, 415–423. doi:10.1016/S1364-6613(03)00197-9

Damasio, A. R., and Marg, E. (1995). Descartes' error: emotion, reason, and the human brain. *Optom. Vis. Sci.* 72, 847–847. doi:10.1097/00006324-199511000-00013

Daniel, T. A., Katz, J. S., and Robinson, J. L. (2016). Delayed match-to-sample in working memory: a brainmap meta-analysis. *Biol. Psychol.* 120, 10–20. doi:10.1016/j.biopsycho.2016.07.015

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi:10.1146/annurev.ne.18.030195.001205

Diamond, A. (1998). Understanding the A-not-B error: working memory vs. reinforced response, or active trace vs. latent trace. *Dev. Sci.* 1, 185–189. doi:10.1111/1467-7687.00022

Donner, T. H., and Nieuwenhuis, S. (2013). Brain-wide gain modulation: the rich get richer. *Nat. Neurosci.* 16, 989–990. doi:10.1038/nn.3471

Doya, K. (2002). Metalearning and neuromodulation. *Neural Netw.* 15, 495–506. doi:10.1016/S0893-6080(02)00044-8

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

Ferrera, V. P., Rudolph, K. K., and Maunsell, J. (1994). Responses of neurons in the parietal and temporal visual pathways during a motion task. *J. Neurosci.* 14, 6171–6186.

Fuster, J. M. (2009). Cortex and memory: emergence of a new paradigm. *Cortex* 21, 2047–2072. doi:10.1162/jocn.2009.21280

Gaffan, D., and Weiskrantz, L. (1980). Recency effects and lesion effects in delayed non-matching to randomly baited samples by monkeys. *Brain Res.* 196, 373–386. doi:10.1016/0006-8993(80)90402-3

Gärdenfors, P. (2003). *How Homo Became Sapiens: On the Evolution of Thinking*. Oxford: Oxford University Press.

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

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., and Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature* 410, 930. doi:10.1038/35073582

Glenberg, A. M. (1997). What memory is for. *Behav. Brain Sci.* 20, 1–19. doi:10.1017/S0140525X97470012

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi:10.1016/0166-2236(92)90344-8

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural Turing Machines. *arXiv preprint arXiv:1410.5401*.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476. doi:10.1038/nature20101

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cogn. Sci.* 11, 23–63. doi:10.1111/j.1551-6708.1987.tb00862.x

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.

Herrmann, M., Ruppin, E., and Usher, M. (1993). A neural model of the dynamic activation of memory. *Biol. Cybern.* 68, 455–463. doi:10.1007/BF00198778

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi:10.1073/pnas.79.8.2554

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. U.S.A.* 81, 3088–3092. doi:10.1073/pnas.81.10.3088

Humphrey, N. K. (1992). *A History of the Mind*. New York: Simon & Schuster.

Jeannerod, M. (1994). The representing brain: neural correlates of motor intention and imagery. *Behav. Brain Sci.* 17, 187–202. doi:10.1017/S0140525X00034026

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Psychol. Sci.* 18, 312–313. doi:10.1111/j.1467-9280.2007.01895.x

Kanizsa, G. (1976). Subjective contours. *Sci. Am.* 234, 48–52. doi:10.1038/scientificamerican0476-48

Kaplan, G. B., Şengör, N. S., Gürvit, H., Genç, İ., and Güzeliş, C. (2006). A composite neural network model for perseveration and distractibility in the Wisconsin card sorting test. *Neural Netw.* 19, 375–387. doi:10.1016/j.neunet.2005.08.015

Lara, A. H., and Wallis, J. D. (2015). The role of prefrontal cortex in working memory: a mini review. *Front. Syst. Neurosci.* 9:173. doi:10.3389/fnsys.2015.00173

Lerner, I., Bentin, S., and Shriki, O. (2010). Automatic and controlled processes in semantic priming: an attractor neural network model with latching dynamics. *Proc. Cogn. Sci. Soc.* 32, 1112–1117.

Lerner, I., Bentin, S., and Shriki, O. (2012). Spreading activation in an attractor network with latching dynamics: automatic semantic priming revisited. *Cogn. Sci.* 36, 1339–1382. doi:10.1111/cogs.12007

Lerner, I., Bentin, S., and Shriki, O. (2014). Integrating the automatic and the controlled: strategies in semantic priming in an attractor network with latching dynamics. *Cogn. Sci.* 38, 1562–1603. doi:10.1111/cogs.12133

Lerner, I., and Shriki, O. (2014). Internally-and externally-driven network transitions as a basis for automatic and strategic processes in semantic priming: theory and experimental validation. *Front. Psychol.* 5:314. doi:10.3389/fpsyg.2014.00314

Lüders, B., Schläger, M., Korach, A., and Risi, S. (2017). "Continual and one-shot learning through neural networks with dynamic external memory," in *European Conference on the Applications of Evolutionary Computation* (Berlin: Springer), 886–901.

Månsson, J. (2006). *Perceptual Surface Reconstruction*, Vol. 129. Lund: Lund University Cognitive Studies.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., and Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science* 315, 393–395. doi:10.1126/science.1131295

Mikami, A., and Kubota, K. (1980). Inferotemporal neuron activities and color discrimination with delay. *Brain Res.* 182, 65–78. doi:10.1016/0006-8993(80)90830-6

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202. doi:10.1146/annurev.neuro.24.1.167

Mishkin, M., and Manning, F. J. (1978). Non-spatial memory after selective prefrontal lesions in monkeys. *Brain Res.* 143, 313–323. doi:10.1016/0006-8993(78)90571-1

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

Muenzinger, K. F. (1938). Vicarious trial and error at a point of choice: I. A general survey of its relation to learning efficiency. *Pedagog. Semin. J. Genet. Psychol.* 53, 75–86. doi:10.1080/08856559.1938.10533799

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a PDP model of the AB task. *Dev. Sci.* 1, 161–184. doi:10.1111/1467-7687.00021

Murdock, B. B. Jr. (1962). The serial position effect of free recall. *J. Exp. Psychol.* 64, 482. doi:10.1037/h0045106

O'Reilly, R. C., Braver, T. S., and Cohen, J. D. (1999). "A biologically-based computational model of working memory," in *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, eds A. Miyake and P. Shah (New York: Cambridge University Press), 375–411.

Page, M., and Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychol. Rev.* 105, 761. doi:10.1037/0033-295X.105.4.761-781

Parisotto, E., and Salakhutdinov, R. (2017). Neural Map: Structured Memory for Deep Reinforcement Learning. *arXiv preprint arXiv:1702.08360*.

Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic Books.

Ponzi, A. (2008). Dynamical model of salience gated working memory, action selection and reinforcement based on basal ganglia and dopamine feedback. *Neural Netw.* 21, 322–330. doi:10.1016/j.neunet.2007.12.040

Redish, A. D. (2016). Vicarious trial and error. *Nat. Rev. Neurosci.* 17, 147. doi:10.1038/nrn.2015.30

Rodriguez, J. S., and Paule, M. G. (2009). "Chapter 12. Working memory delayed response tasks in monkeys," in *Methods of Behavior Analysis in Neuroscience*, ed. J. J. Buccafusco (Boca Raton, FL: CRC Press/Taylor & Francis).

Russo, E., Namboodiri, V. M., Treves, A., and Kropff, E. (2008). Free association transitions in models of cortical latching dynamics. *New J. Phys.* 10, 015008. doi:10.1088/1367-2630/10/1/015008

Russo, E., and Treves, A. (2012). Cortical free-association dynamics: distinct phases of a latching network. *Phys. Rev. E Stat. Nonlin. Soft Matter. Phys.* 85, 051920. doi:10.1103/PhysRevE.85.051920

Sawaguchi, T., and Yamane, I. (1999). Properties of delay-period neuronal activity in the monkey dorsolateral prefrontal cortex during a spatial delayed matching-to-sample task. *J. Neurophysiol.* 82, 2070–2080. doi:10.1152/jn.1999.82.5.2070

Smith, M. L., and Milner, B. (1981). The role of the right hippocampus in the recall of spatial location. *Neuropsychologia* 19, 781–793. doi:10.1016/0028-3932(81)90090-7

Sompolinsky, H., and Kanter, I. (1986). Temporal association in asymmetric neural networks. *Phys. Rev. Lett.* 57, 2861. doi:10.1103/PhysRevLett.57.2861

Sylvester, J., and Reggia, J. (2016). Engineering neural systems for high-level problem solving. *Neural Netw.* 79, 37–52. doi:10.1016/j.neunet.2016.03.006

Sylvester, J., Reggia, J., Weems, S., and Bunting, M. (2013). Controlling working memory with learned instructions. *Neural Netw.* 41, 23–38. doi:10.1016/j.neunet.2013.01.010

Tanji, J., and Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Curr. Opin. Neurobiol.* 11, 164–170. doi:10.1016/S0959-4388(00)00192-6

Tolman, E. C. (1939). Prediction of vicarious trial and error by means of the schematic sowbug. *Psychol. Rev.* 46, 318. doi:10.1037/h0057054

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189. doi:10.1037/h0061626

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

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. *Neural Netw.* 10, 821–835.

Verduzco-Flores, S., Ermentrout, B., and Bodner, M. (2012). Modeling neuropathologies as disruption of normal sequence generation in working memory networks. *Neural Netw.* 27, 21–31. doi:10.1016/j.neunet.2011.09.007

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* 24, 455–463. doi:10.1016/S0166-2236(00)01868-3

Weston, J., Chopra, S., and Bordes, A. (2014). Memory Networks. *arXiv preprint arXiv:1410.3916*.

Conflict of Interest Statement: 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.

HYPOTHESIS AND THEORY
published: 18 April 2018
doi: 10.3389/frobt.2018.00039

Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy

Riccardo Manzotti[1] *and Antonio Chella*[2,3]*

[1] Department of Business, Law, Economics and Consumer Behavior, Università di Comunicazione e Lingue (IULM), Milan, Italy, [2] RoboticsLab, Department of Industrial and Digital Innovation, University of Palermo, Palermo, Italy, [3] Cognitive Robotics and Social Sensing Laboratory, ICAR-CNR, Palermo, Italy

Recently, there has been considerable interest and effort to the possibility to design and implement conscious robots, i.e., the chance that robots may have subjective experiences. Typical approaches as the global workspace, information integration, enaction, cognitive mechanisms, embodiment, i.e., the Good Old-Fashioned Artificial Consciousness, henceforth, GOFAC, share the same conceptual framework. In this paper, we discuss GOFAC's basic tenets and their implication for AI and Robotics. In particular, we point out the intermediate level fallacy as the central issue affecting GOFAC. Finally, we outline a possible alternative conceptual framework toward robot consciousness.

Keywords: robot consciousness, machine consciousness, artificial consciousness, synthetic phenomenology, robot self-awareness

OPEN ACCESS

Edited by:
Nadia Magnenat Thalmann,
Université de Genève, Switzerland

Reviewed by:
Leonid Perlovsky,
Harvard University, United States
Alessandro Di Nuovo,
Sheffield Hallam University,
United Kingdom

****Correspondence:***
Antonio Chella
antonio.chella@unipa.it

Specialty section:
This article was submitted to
Humanoid Robotics,
a section of the journal
Frontiers in Robotics and AI

Received: *30 November 2017*
Accepted: *20 March 2018*
Published: *18 April 2018*

Citation:
Manzotti R and Chella A (2018) Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy. Front. Robot. AI 5:39.
doi: 10.3389/frobt.2018.00039

INTRODUCTION

Consciousness exists: we are conscious, and it would be odd to negate this fact. Consciousness is a part of our physical world, and then the processes at the basis of consciousness must be faced by the laws of science governing our physical world.

The definition of consciousness is still an open question. Therefore, it would be problematic to discuss about robot consciousness: in facts, Raoult and Yampolskiy (2015) reviewed 21 proposed tests presented in the literature to assess consciousness in machines and robots. However, the same situation holds for other complex concepts: notably, Legg and Hutter (2007) review more than 70 existing different definitions of "intelligence." The fact that there is no agreement on what intelligence is does not refrain researchers to speaking about Artificial Intelligence.

In facts, consciousness is an important research topic in neuroscience: Dehaene (2014) summarizes several years of studies in human consciousness; see also Tononi (2012) and Damasio (2010), among others. Notably, neuroscientists working on consciousness take seriously into account the possibility that, in the near future, robots may be conscious. During the Symposium organized in 2001 by the Swartz Foundation on "Can a Machine Be Conscious," the concluding remarks of Christof Koch stated that:

> "we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artifacts designed or evolved by humans.[1]"

To the best of our knowledge, this claim is valid still today.

Consciousness is part of our physical world, and then some of its aspects may be studied and even replicated by using robots. On the one hand, the employment of robots as tools may help to

[1] http://www.theswartzfoundation.org/abstracts/2001_summary.asp

understand biological consciousness better, and, on the other hand, the processes at the basis of consciousness may be in some sense crudely replicated to build better robots, as it happened, e.g., for neural networks and artificial life systems. Anil Seth has claimed that:

> "Over the last *two* decades much has changed [...]. Alongside philosophical discourse a new science of consciousness has taken shape which integrates experimental and theoretical work cross many fields including neuroscience, psychology, cognitive science, artificial intelligence, computer science, neurology, and psychiatry." (Seth, 2010, p. 1).

It is not a case that the late Nobel prize Gerald Edelman, a scholar of the research on consciousness, employed robots to validate parts of his theory of consciousness (Reeke et al., 1990; Edelman et al., 1992). Koch and Tononi directly addressed the possibility that artifacts may be conscious by taking into account constraints and conditions according to the Integrated Information Theory of consciousness (see below). Notably, Koch and Tononi (2008, 2017) explicitly discussed and proposed tests for consciousness in the machines. Recently, Dehaene et al. (2017) summarized the neuroscientific findings of interest for conscious machines. We concur with their claim according to which the study of biological consciousness may inspire novel machine architectures.

In this direction, the paper by Grossberg (2017) summarizes years of works about brain resonances and proposes a set of models, described by differential equations, that captures some of the main aspects of consciousness. Important papers in this line are due to Perlovsky (2006, 2016) where he claims that a new "physics of mind" is needed that looks for the fundamental laws of the material world, including sentience. The new physics of mind should develop the mathematical theories that explain the empirical evidence about sentience and that generate suitable predictions to be verified by experiments.

Therefore, the problem of consciousness in robots and artifacts is an accepted issue for researchers in neuroscience.

In the AI debate, the problem of machine consciousness has been discussed by many scholars since the seminal paper by McCarthy (2002), where he considered an extension of the Situation Calculus to deal with some aspects of self-reflection to make robot conscious of their mental states. On a similar line, McDermott (2001) devoted his book on "Mind and Mechanisms" on the discussion of a computational theory of consciousness.

Many journals and conference papers discussed the possibility of consciousness in machines and robots by proposing theories and architectures. Holland (2003) and Chella and Manzotti (2007) collected the initial attempts at robot consciousness. An almost complete up to date review is due to Reggia (2013). Scheutz (2014) reviewed and discussed the contact points between machine consciousness and artificial emotions.

Among the essential works from AI scholars concerning machine and robot consciousness, we mention, among others, the architectures based on the global workspace model of consciousness (Baars, 1997) as the LIDA architecture (Franklin, 2003; Franklin et al., 2014) and the cognitive architecture proposed by Shanahan (2005, 2006). A model of conscious experience related to learning and sensorimotor interaction in an autonomous robot has been discussed by Kuipers (2008). Notably, Bringsjord et al. (2015) recently implemented a cognitive system based on higher-order logic running on the NAO robot that passed human tests of self-consciousness.

Therefore, robot consciousness is an important research field that benefits from the contributions of many scholars from neuroscience, Artificial Intelligence, and robotics. The general feeling is, as stated above, that understanding biological consciousness may help to build better robots and, on the other side, that the research on robot consciousness may help understanding biological consciousness.

This paper aims to propose a critical review and analysis of the literature related to robots and machine consciousness under the light of what we named the "intermediate level fallacy." In facts, many theories of machine consciousness actually do not directly address the problem of consciousness, but they discuss some intermediate problem, then leaving aside the issue of robot consciousness.

Then, the goal of the paper is not to discuss a specific algorithm or software, but to help roboticists interested in robot consciousness to build a mental map of the bibliography in the field and to avoid quirks due to the intermediate level fallacy.

GOOD OLD FASHIONED MACHINE CONSCIOUSNESS

As previously stated, in recent years the notion of machine and robot consciousness gained momentum and attracted considerable interest. Chella and Manzotti (2009) discussed many problems arising in the assessment of consciousness in a robot concerning the role of the body, the needs for the robot to be "situated" in an environment, the cognitive capabilities of the robot, the effective functions of emotions and so on.

The most challenging problem for robot consciousness is the possibility that a robot may have real subjective experiences. However, many approaches at the state of the art in robot consciousness are biased by a set of premises that harnessed research into what can be named as Good Old-Fashioned Artificial Consciousness (GOFAC).

GOFAC suggests a physical world in which consciousness appears as a result of a specific intermediate level. A theory based on the idea that consciousness emerges from an intermediate level should explain what this level is and why it produces consciousness. However, the explanation is problematic because, rather than explaining consciousness, the theory introduces a new level as an intermediate entity, that is only apparently less troublesome. In contrast, the intermediate level is explanatory disruptive since it adds two new problems: the characteristics of the new level and its relation to consciousness. This approach can be named the "intermediate level fallacy" and it seems to be attractive because the introduced level appears less intimidating and more familiar than consciousness itself.

This paper aims to list some of the leading approaches to robot consciousness under the light of the intermediate level fallacy. While each method has its peculiar shortcomings, they share the standard pattern, i.e., the intermediate level fallacy, that characterizes GOFAC. They try to downgrade the notion of consciousness to something more amenable—a move that does not solve, but it multiplies the problems. After analyzing a series of well-known approaches, the paper outlines a possible direction in which research might go to overcome GOFAC.

This paper has not a negative goal, namely to list a series of hypotheses and premises and stress the overall failure of GOFAC. Instead, it aims to shed light on some, likely fruitful direction of research.

THE HARD PROBLEM

The main culprit behind GOFAC is David Chalmers's introduction of the *hard problem* (Chalmers, 1996). According to Chalmers's seminal book, most of research and discussion about consciousness has been carried on inside the conceptual framework set by the contrast between a *conscious* mind and a *cognitive* mind. Such a notion has entrenched the gap between subjective, *phenomenal* experience and *physical* properties. The hard problem—namely the idea that once all the material facts are fixed, there is still something to be explained, has postponed the understanding of consciousness and placed it outside of robot implementation. If one accepts it, it follows that a robot will never be genuinely conscious because no matter how all physical facts are fixed, there will still be something to be added. The acceptance of the hard problem is the main reason behind the ensuing lack of progress in robot consciousness.

The hard problem is based on the premise that subjective and physical properties are alien to each other. Moreover, yet, this premise is not of experimental nature, and it might be questioned. In facts, if subjective and physical properties are different, then it would be impossible to place them against each. Consider, for example, the comparison between *subjective red* and *real red*. There is no reason to believe there are two kinds of red. Of course, the usual claim is that the subjective red is of mental nature and the physical world is not accessible in a profound sense. Chalmers claims that there are only subjective properties, or, to use an equivalent and famous formulation, that we only experience the phenomenal character of what happens.

This claim is unsupported by the facts that human beings experience the external world and their own body. It is not phenomenal; it is just what the physical world is. There are no reasons to assume, as Chalmers does, that the perception of the world is different from the physical world. There are no perceptions of subjective properties, but instead, human beings experience the attributes the world is made of, and the name we can give to such characteristics is physical.

The hard problem is not empirically grounded because if it were true, it could not be empirically proven. If consciousness were hard, it could not affect the physical world. Conversely, if consciousness were testable, it would not be hard.

The hard problem is related to the *epiphenomenal* conception of consciousness, i.e., that consciousness has no physical role. Accepting the hard problem means that consciousness will be external to the domain of material facts. In fact, if consciousness were part of the physical world, it could be measured, observed, replicated, designed and implemented in a robot. The hard problem encourages to conceive of consciousness as something intractable by scientific means. Consciousness could not have any effect on the physical world and, consequently, it would be useless from a robotics perspective.

However, if consciousness is epiphenomenal, it would contradict the selective advantage that it seems to provide. Moreover, there are no other natural phenomena that are deemed to be epiphenomenal. All physical events are causally relevant, that is why they can be measured and observed them, as they exert a causal effect. In physics and engineering, there are no such phenomena because they would be automatically deemed not to be real. The fact that GOFAC deals with consciousness as epiphenomenal is the hallmark of scientific failure. Once inside the traditional GOFAC framework, then consciousness is outside of empirical reach.

The notion of epiphenomenal consciousness appears to be a self-defeating hypothesis. Human beings as conscious agents have a feeling that what they feel is interwoven with the physical world. Consciousness is indeed a part of the physical world, and if the current scientific picture of the world does not have a place for consciousness, then it is not complete.

Nonetheless, the hard problem became famous also because it contrasts the *easy* problems—how to explain the human ability to recognize a face, generate language, control behavior—from the hard problem of defining how physical processes can give rise to consciousness. Such a split suggests, on the one side, that scientists could continue their work without worrying about consciousness and, on the other hand, that consciousness is elusive and not constrained by the physical world. It also provides engineers, roboticists, and AI experts free to design robot consciousness as long as they were smart enough to leave the hard problem aside and limit themselves to the easy problems of consciousness.

In GOFAC, the hard problem spawned a split between hard and weak machine consciousness (Seth, 2009) as though it were possible to focus on functional and ontological problems separately. Because of the widespread acceptance of the hard problem, scholars assumed that conscious experience is out of reach of science and technology and thus that a workaround has to be proposed. The workaround was the delusion that it is possible to focus on concrete problems—i.e., those that are part of our conceptual framework—and to leave the real issue of consciousness to some conceptual breakthrough.

The above state of things suggests that the literature on robot consciousness does not deal with phenomenal consciousness. Consciousness has been dropped from the physical world by the hard problem, and thus it has been become legitimate to study it without addressing the crux of the matter.

THE INTERMEDIATE LEVEL FALLACY

Given the starting conceptual landscape shaped by the acceptance of the hard problem—or some version of it—a widespread tendency has been that of looking for some workaround. A common strategy has been that of the intermediate level which is composed of two steps. First, an intermediate conceptual level that is at a possible explanatory distance is proposed—behavior, central workspace, information, enaction, adaptive resonance, and so forth. Such an entity, crucially, is located on the physical side of the gap but, equally significantly, it is somewhat vague, to the extent that it may suggest some degrees of consciousness. Second, consciousness is watered down to show that it is not much better than the intermediate level. The second step, which is most problematic from an ontological and epistemic perspective, is critical to provide fulfillment of the first step.

As an example of the intermediate level fallacy, consider Seth's proposal to look for a real problem rather than for the hard or the easy problem. According to Seth, the real question consists in examining

> "how to account for the various properties of consciousness regarding biological mechanisms; without pretending it doesn't exist (easy problem) and without worrying too much about explaining its existence in the first place (hard problem)." (Seth, 2016).

The real problem, according to Seth, is nothing but one of the traditional easy problems in disguise. In this case, the intermediate level is represented by the biological mechanisms that are physical processes that do not qualify as a solution to the hard problem. In this regard, Seth himself defended weak machine consciousness (Seth, 2009). So, it is not clear why the real problem according to Seth should be a successful research strategy for consciousness. It is the second step of the intermediate level strategy, i.e., watering down consciousness. Seth's catchphrase is that

> "It looks like scientists and philosophers might have made consciousness far more mysterious than it needs to be" (Seth, 2016).

Thus, he suggests that, after all, there is no mystery. In fact, Seth argues that

> "In the same way, tackling the real problem of consciousness depends on distinguishing different aspects of consciousness, and mapping their phenomenological properties (subjective first-person descriptions of what conscious experiences are like) onto underlying biological mechanisms (objective third-person descriptions)" (Seth, 2016).

In his account, the problem of consciousness is no longer that of tackling an apparently impossible feat for the physical world, but a mapping between personal reports onto biological mechanisms. This mapping may be tedious but feasible. However, such a mapping does not offer a solution of the problem of consciousness. Both personal descriptions and biological mechanisms are objective physical phenomena that pose no threat to the received view of physics. Both of them do not address the issue of consciousness.

Then, the first step of the fallacy is to suggest an intermediate, safe level of explanation, like a suitable biological mechanisms. The second step is to water down the problem of consciousness to something more amenable as the mapping between personal reports and the biological mechanisms.

CURRENT APPROACHES TO ROBOT CONSCIOUSNESS

Robot consciousness has so far not succeeded in making progress on the issue of phenomenal experience. While the possibility of conscious machines, together with its ethical implications, has repeatedly been addressed, no one has claimed that anything close to a feeling has occurred in an artifact. As before, this persistent and generalized lack of results might be explained by the adoption of the familiar and flawed conceptual landscape of GOFAC. In particular, the intermediate level fallacy is a common problem in all these attempts. Here, we will consider, as possible theoretical backgrounds for machine consciousness, functionalism, information, embodiment, enaction and cognition. We will argue that these approaches exhibit the manifest symptoms of the fallacy and are as many cases of GOFAC.

Functionalism

Functionalism is the backbone of the AI approach to consciousness. Functionalist approaches single out a functional view of the mind. This critique has been developed at length by many scholars, most notably Searle (1990) and Harnad (2003). If the mind is a collection of functional relations, no space is left for what is taken to be consciousness—functioning vs. feeling, to use Harnad's formulation. Functionalism focuses on external causal relations between the state of affairs. While functionalism is neutral to the location of such causal relationships, it concentrates mostly on abstract descriptions of reality, which is the reason why it allows multiple realizations. Functionalism is a theoretical description of what goes on in a system, and it is oblivious to the physical constituents of a system. Therefore, functionalism will never grasp consciousness because it is neutral to the material components of functional relations.

Then, functionalism would provide the same description for a system made of neurons and of electronic switches, and it will offer the same explanation for a system with consciousness and without consciousness. It is not a fact about consciousness; it is a consequence of the premises on which functionalism is built.

Functionalism has been ideal to back up the philosophical notion of a *zombie*, which was fundamental in all the accounts inspired by the hard problem (Chalmers, 1996). A zombie is an entity which externally is not distinguishable from a human being, in the sense that it talks, it responds, it acts in the world, but, contrary to a human being, it is entirely unconscious. The conceivability of a zombie tells us more about the limitation of functionalism than about consciousness. There is no evidence that a physical entity identical to a human being might be without consciousness. The notion of a zombie shows that functional descriptions are incomplete and leave out something crucial. In fact, in practice, all machines nowadays are considered philosophical zombies. No one expects Siri or Google Assistant to be anything but zombies.

Many approaches to consciousness are functionalist models. Consider the mentioned global workspace model (Baars, 1997) and its implementations (Shanahan, 2005, 2006; Franklin et al., 2014). Such a model is constituted by a suitable functional structure where the information is lumped and broadcasted. The first step is represented by the particular cognitive structure, the central workspace, that is a neutral concept, and that takes into account the notion of unity and the idea of a central controller. The second step is the watering down of consciousness, namely the claim that, to be conscious is nothing but accessing information in a centralized fashion.

Another approach is the model of consciousness formulated by Stephen Grossberg (2007, 2017) and based on adaptive resonances in the brain. According to this model, the conscious states in the brain are characterized as resonant neural states, i.e., neural states where the firing of neurons are mutually amplified and synchronized thanks to feed-forward and feedback connections between bottom-up and top-down neural layers. In this case, the first step of the move is represented by a suitable characteristic of the dynamic evolution of a neural network, i.e., the resonance of interconnected neurons, which is a neutral effect that is explained by the differential equations governing the dynamics of neural networks. The second step is the claim that subjective experience is nothing but this particular state in the dynamic evolution of neural networks. Of course, not any rationale has been presented as to why centralized accessed information or a resonant state could not be unconscious. The presence of the fallacy is evident.

It is not to say that robots envisaged by functionalist designers will never be conscious. In fact, designers, no matter what conceptual frameworks they employ, when they move from designing to implementations, are subject to the structure of the physical world. Thus, their products are not limited by their conceptual models. As consciousness is part of the natural manifold, there will be cases in which the physical structure of agents will yield to consciousness, no matter the conceptual framework adopted by its designers.

Information and Computation

Another popular approach in GOFAC is based on seeking unique information processes that produce consciousness. Information, at the level of computational processes such as those implemented by brains or by computers, is not a physical constituent of reality. Instead, it is a convenient level of description. Information is a fictitious entity, like a center of mass or a meridian: it is not physically there, but it exists only in our descriptions. It cannot be observed, but, significantly, calculated.

In the case of information, there is confusion among scientists. The everyday familiarity with information has fostered a widespread tendency to deal with information as though it were real, like water or electricity. However, there is no evidence that information is anything over and above the physical processes we describe using an informational jargon (Shannon, 1948; Searle, 1984); it is nothing but a quantitative description of the causal relations between events. From a physical perspective, there is no need for an additional level called information over and above the physical phenomena, but all the causal power is drained by physical events (Kim, 1989, 1998; Dowe, 2000, 2007).

As an argument of the fact that information does not have a physical existence consider that if information were real, it should be possible to build an information detector. Interestingly, it is not possible to construct an information detector. While it is possible to compute the amount of information inside a system from a set of assumption as to how that system is going to be exploited, it is impossible to detect the amount of information in a system. For instance, if one knows that a CD-Rom is going to be read by a standard CD-Player one can compute its capacity. However, if one takes a piece of matter and one does not know whether and how its physical structure is going to be exploited, one cannot know how much information it contains. The same holds in all cases of similar information devices. It is not possible to measure information as say, mass, electric charge, length. Information can be *estimated* or *computed* based on what it is known about a piece of matter and its role in a given context.

In sum, information does not exist except as a way to describe what does happen between causally coupled events, coherently with the original formulation of information (Shannon, 1948). Information is a way to explain causal processes; it is not a real phenomenon. It is not physical insofar it is causally redundant, undetectable, never measured but only estimated. On top of that, there would be no law explaining why a specific informational state should be like a conscious state.

Information-based approaches to consciousness remain in the intermediate level fallacy. The intermediate entity is now information—sometimes a specific brand of information as in Tononi's integrated information theory (Tononi, 2004) and its most recent version (Oizumi et al., 2014). The watering down is the effort to claim that the properties of information are those that matter for consciousness. For instance, Tononi claimed that integrated information has unity and that consciousness too has unity. Concerning quality, semantics, content, and all other aspects of our experience, he does not have any word.

In sum, approaches like those suggested by Tononi and based on the idea that information processing produces consciousness, are empirically not founded because information has not a physical reality. They are biased by the hope that a quantitative, precise method may offer a scientific framework. In fact, these authors emphasize the possibility to *measure* consciousness. At most, these methods can succeed in estimating informational states that correlate with consciousness, but, so far, they have been unable to present justification as to why the informational states under scrutiny should constitute consciousness.

Embodiment

In robot consciousness, popular approaches are related with the notion of embodiment (Holland, 2004; Bongard et al., 2006; Shanahan, 2006, 2010) mostly because they allow focusing on robot bodies. It is a fruitful approach that highlights crucial features of the embodiment. The body plays an essential role in shaping the interaction between an agent and its environment. Embodied cognition is a mandatory perspective regarding sensory-motor loops. However, it is not clear why embodiment should provide clues on how consciousness fits with the physical world. Inevitably, embodiment simplifies many critical sensory-motor control loops.

If embodiment refers to the fact that a cognitive or conscious process must be physically embodied, it is a pretty obvious notion. A cognitive process must be embodied in this sense, as any process must correspond to something physical and thus be embodied. However, supporters of the concept of embodiment as Chrisley and Ziemke (2006) mean something less trivial.

These authors compete against the traditional notion of cognition as a higher order process carried on by a central processing unit physically separate from the body. Such an approach is the offshoot of historical factors—i.e., mostly, the Cartesian notion of an immaterial mind, a functionalist model of the mind, and the availability of electronic calculators well before they could be coupled with artificial bodies. All these factors fostered a disembodied notion of the mind and its processes. However, they have long ceased to be relevant, both in the philosophical debate as well as in the technological playground.

AI is biased by a Cartesian view of the mind. Embodiment allowed AI scholars to emphasize the physical nature of agent hood. However, this fact does not imply that the body is the only constituent of an agent.

The notion of embodiment self-contradicts its original intentions. In fact, the embodiment was taken into consideration to get rid of the immaterial mind, as the body and its interaction with the world appear like a feasible solution. Unfortunately, the notion of "body" is unclear. Typically, an object is a body only when it is the body of a subject. However, then, the notion of the body is circularly the cornerstone of the subject. The body is another intermediate entity that should bridge the gap between world and consciousness. It is the symptom of the intermediate level fallacy. The body—or its interactions with the environment—is proposed as the intermediate level. At the same time, the watering down step deals with the body as though it were something more than a moving physical object. The last step is, of course, of relevance in the case of robot consciousness where researchers do not have a biological body. The features that should be present in an object to be qualified as a body are not explained. In this sense, a washing machine may be considered as a body, because it reacts to external stimuli, it swallows stuff, it processes it, it expels it, it consumes energy, it plans. The same arguments hold for anthropomorphic robots (Holland, 2003; Natale et al., 2012).

Thus, embodiment tries to exploit the intermediate level fallacy by employing the ambiguous notion of a body, and to water down consciousness to something more mundane as the body.

Enaction

Another viable solution to achieve robot consciousness is offered by enaction insofar as it suggests that experience is constituted by a body and its interactions and with the world, and thus it may be implemented in artifacts (O'Regan and Nöe, 2001).

Enactivism defends a firm stance that, together with the embodiment is likely to be productive in many fields, most notably cognitive science (Stewart et al., 2010). What enaction has never addressed is the enactive level of reality and why there should be anything like that—namely the first step of the fallacy.

Consider the basic tenet of enaction, in Alva Noë's formulation:

> "Perceiving is a way of acting [. . .] What we perceive is determined by what we are ready to do [. . .] We enact out perception; we act it out" (Noë, 2004, p. 1).

Once again, Noë suggests an intermediate level based on actions, that should underpin perception. Of course, he does not explain why actions should be different in the case they are performed by human bodies from the case in which they are performed by a robot or an animal.

Enactivism does not provide a criterion to distinguish between real actions and simple movements unless by reference to subjects. In other words, an act is a movement performed by a subject with intentions and understanding—i.e., a conscious subject. Then there is the concrete risk of circularity in their arguments. Consider this point in John Stewart's formulation:

> "How can a material state *be* a mental state? Hoary it may be, yet the problem is anything but solved. [. . .] The paradigm of enaction solves this problem by grounding all cognition as an essential feature of living organism" (Stewart, 2010, p. 1).

Of course, as Stewart himself admits, this does not solve the problem. It only shifts the burden of the explanation on the notion of the living organism. Since vitalism has long been dismissed, the emphasis on life and living organisms does not seem a convincing conceptual fulcrum. In this way, the suggested intermediate level is the living organism and its feedback loops with the external world. Why these phenomena should be any special is left unexplained. It is the second step of the fallacy.

Finally, it is characteristic of enaction the shift from actions as such to knowledge about actions. In fact, recent accounts of consciousness in enaction take stock of the notion of knowledge. In this regard, Noë claims that

> "To be a perceiver is to understand, implicitly, the effects of movement on sensory stimulation." (Noë, 2004, p. 1).

Once again, an intermediate level, that of understanding and sensory-motor knowledge, is presented as a way to reach consciousness. What such an intermediate level is in a physical world and why knowledge of the effects of movement on sensory stimulation should lead to conscious experience is not clear at all.

Cognition and Intelligence

The most obvious candidate for consciousness is cognition and intelligence. Here, we have a promising intermediate entity which looks apparently less demanding, and we may consider whether it might be the right ladder. After all, there seems to be a tight connection between cognitive capabilities and consciousness. Most of the time, when a human being exerts higher-order cognitive processes are conscious. However, it is fair to maintain that, in many cases, when one is conscious very little intelligence is required or that many of the most creative ideas have been the outcome of mostly unconscious activities (Lavazza and Manzotti, 2013).

It is a fact that many scholars are tempted to focus on intelligence and cognition and expect that consciousness will come for free once all the practical issues have been solved. Alternatively, instead many hold that the problem will evaporate as a false problem.

However, also, in this case, cognition is an intermediate level that may lead to the knowledge of consciousness, and not to consciousness experience. Also, this is a symptom of the intermediate level fallacy.

WHAT IS LEFT?

We found a common explanatory strategy in the reviewed attempts. Scholars working in robot consciousness suggest an intermediate level—sensory-motor patterns, information, cognition, global workspace—as a possible explanation for consciousness. What is missing is why such a level should lead to consciousness. From an epistemic perspective, it is as though they suggested an *explanans* without providing its relationship with the *explanandum,* i.e., consciousness. **Table 1** summarizes the different GOFAC landscapes of the intermediate level fallacy.

The hard problem, the GOFAC approaches, and the strong vs. weak machine consciousness argument are all grouped by a common factor, as they all deal with consciousness as lacking any causal role in the world. Consider for example the hard problem, that leads to the issue of the zombie, a cognitively equivalent agent lacking consciousness. In turn, GOFAC does not address the issue of subjective experience. Finally, the split between weak and strong machine consciousness was conceived to deal with cognitive processing without addressing the crux of the matter, namely conscious experience. Weak consciousness was designed to deal with the functional aspects of consciousness—i.e., those with causal relevance—and therefore to leave out strong consciousness.

New hypotheses about the nature of the physical world are needed. Consciousness is a fact that needs to find its place in nature. Thus, if consciousness is neither of the previously examined processes what is left? The proposal is that consciousness is the structure of the physical world itself. Such a move has been except in some cases, as in Perlovsky (2006, 2016). There must be fundamental mistakes in the way the physical world is conceived. A possible error might be the location of the thing called consciousness in a different place rather than the body of the agent or the neural/computational structure. Another mistake might consist in the split between the subject and the object. The paper shows that GOFAC will never achieve machine consciousness and thus that it clamors for the adoption of a robust conceptual framework alternative to the hard problem and its cognates.

Of course, finding consciousness inside the physical world is necessary when the goal is designing a conscious robot. A robot does not have any other resource but those offered by the physical world. It may sound like a platitude but, give or take, all mentioned approaches run according to this principle. Therefore, any viable solutions will require setting aside the premise that has so far hampered any progress—i.e., the hard problem with the general belief that consciousness is something distinct from the physical world. We have to reconsider the question from the beginning.

We believe it is possible to flesh out a radical alternative that will stem from setting aside the obnoxious theoretical framework fostered by the adoption of the Hard Problem. First, we take consciousness to be just like all other physical properties around, something that can be measured, observed. Furthermore, consciousness is causally active and located in space-time. Finally, it is made of matter or energy. These premises are nothing more than restating the assumption that consciousness is physical. In fact, everything that is physical is spatiotemporally located, causally relevant, made of matter/energy, and observable. So much the worse for epiphenomenalism and zombies.

Of course, this move will be considered unfeasible by most scholars insofar as they take consciousness to be invisible in the physical world. Neuroscientists have been looking for it inside the brain for the last couple of centuries without finding anything resembling it. In the brain, there is nothing like conscious experience and thus neither will there be inside a machine. However, the solution might require a conceptual leap.

Consider the possibility that consciousness, albeit physical, is not literally inside the body of the agent—be it biological or artificial. The proposal is that consciousness is the same with the external objects an agent deals with. In this way, the physical properties of the external world might be the same as

TABLE 1 | The intermediate level fallacy in different GOFAC landscapes.

	Actual physical world	Intermediate level	Watered down version of consciousness
Functionalism	The physical states that realize functional structures	Global workspace, centralized representations, adaptive resonance	Access consciousness
Information and computation	The physical states that transmit causal processes	Integrated Information	Integrated consciousness
Embodiment	Objects	Body states, body-world states	Sensory-motor loops
Enaction	Interactions between objects and environment	Actions	Knowledge of sensory-motor loops
Cognition	Brain or processor	Cognitive states	Knowledge

the properties of conscious experience (Manzotti, 2006, 2017; Manzotti and Chella, 2016).

An example will help. An agent—i.e., a body either biological or a robot—is interacting with an external object, say, a yellow banana. Inside the agent there is nothing with the properties of the banana—being yellow, being elongated, and being slightly bent. When we look for the agent's experience inside the agent's body, we would be compelled to conclude that there is nothing physical with those properties inside the agent's body (yellow, elongated, bent). Not being able to find anything like our experience inside one's body, we may be tempted to conclude that consciousness is indeed particular; that it is invisible, epiphenomenal, not directly measurable, in a world, that it is phenomenal. This option is taken by the hard problem and all its cognate approaches.

We suggest an alternative. When the agent is interacting with the banana, there is a physical entity that is ideally suited to be the same with the agent's experience, namely, the banana itself. The banana is yellow, elongated and slightly bent, just like the experience of it. Nothing else is to be invoked to be the experience of the banana. The banana is better than anything we may ever hope to find inside the agent. The external object scores better than any internal representations.

The advantages of this approach as regards machine consciousness are numerous. There is no need for biological material. There is no need for the emergent property, a very questionable addition to the debate. There is no need to appeal to quantum mechanics, something still alien to the current state of the art in robotics. There is no need to suppose the existence of dubious properties that cannot be observed physically. Everything is measurable, observable and, crucially, causally relevant rather than epiphenomenal. An initial example of this approach, implemented on a robot head, is described in details in Manzotti and Tagliasco (2005).

DISCUSSION

Four possible objections can be anticipated to this proposal. First objection: the object is not inside the body of the agent, and thus it cannot be either constitutive or the cause of one's experience. This objection has been raised by one of the original proposers of the extended mind, namely by Chalmers (2008). There is no reason to assume that we are located in our head. The physical location of experience cannot be derived from the fact that sensor organs are found on the body. Only the location of sense organs can be estimated by the position of what is perceived. The physical location of consciousness is immaterial, though, as Daniel Dennett's clarified in his famous cautionary tale (Dennett, 1978).

Second objection: the yellow of the banana is not like the yellow of consciousness, or to rephrase it, *the physical yellow is different from the phenomenal yellow*. If we assume that subjective properties are different from physical properties, they could not be the same. This fact, however, is neither self-evident nor empirically found. It is the premise on top of which the hard problem framework got built, an assumption that should be empirically demonstrated rather than assumed. In fact, such a hypothesis is self-confuting—if the two classes of properties were different, we could never see the physical properties. The claim that physical properties are different from subjective properties is unproven. The burden of the proof lies on the shoulder of those who claim there are additional properties. Historically, many scholars argued there where subjective properties because they could not find anything like our experience inside brains. However, external objects are exactly like our experience of them. Therefore, nothing prevents from being the same with our alleged experience of them.

Third objection: the misperception as dreams and hallucinations. Any realist proposal must tackle the issue of misperception. How can the suggested identity between consciousness and external object tackles cases in which the object does not seem to be there? Our reply to such an objection is that the scope of the *present* can be arbitrarily large. Consciousness is made of objects that had causal intercourse with the body of the agent and that, thanks to its neural structure, are still causally active in whatever combinations they happen to be. Consciousness is then always a form of perception, albeit reshuffled and postponed. Of course, this issue alone will require a lot more discussion, but the gist of the strategy is there.

Fourth objection: if consciousness is the same with the external objects, how can the same object look different to different agents? A reply is the following—physical properties are relative, and thus they can be different when compared to a different physical system. The same object can have different physical properties for different agents since different agents have different bodies. The same vehicle can have different velocities relative to different observers moving with as many frames of references. So, the same object can have different properties relative to bodies having different causal properties. The same object will have different colors for tetrachromats, standard trichromats, and color blind of various kinds. Thus, the relative nature of physical properties paves the way to the fact that the same object may indeed have different features for different agents.

CONCLUSIONS

The purpose of this article is to show the problems with GOFAC and thus that it clamors for the adoption of a robust conceptual framework alternative to the Hard Problem and its cognates.

Our proposal offers a new basis for robot consciousness. There will no longer be an elusive property concocted by some particular process inside the body of a robot agent; neither will it be a hard problem. Consciousness is the network of objects and events that, thanks to a body with sensory-motor-cognitive capability are brought to interact together. Consciousness is not an internal property, but the collection of objects that, thanks to the body, are causally responsible for what the body does. The study of robot consciousness will thus shift the focus from internal processes and structures to the analysis of the ontogenetic and epigenetic relations that a body develops and maintains with the external world during its life. Methodologies of developmental robotics (Cangelosi and Schlesinger, 2015) will be a valuable help in this effort.

The presented hypothesis, albeit still in its infancy, offers a complete physicalist alternative—conscious robots would be machines that bring into existence the same relative physical objects human bodies do.

The advent of a conscious robot would eventually lead to new questions about what it means to be a person. The concept of person undergone inclusive variations over the centuries, as discussed by Gunkel (2012). Humanity has come across many problems to include women, slaves and superior mammals in the circle of persons. Today, the problem is two-fold: if we assert that a robot is a kind of person, then the moral responsibility of the robot for its actions must be recognized. On the other side, we have to concede some moral rights to the robot, such as the right of not being switched off.

The concept of person is tightly linked to the concept of consciousness. If an entity can have subjective experiences, and eventually can suffer, then this entity should be treated as a person. In this regard, the studies on robot consciousness may force us to review our fundamental definition of the concept of person.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Baars, B. J. (1997). *In the Theather of Consciousness. The Workspace of the Mind.* Oxford: Oxford University Press.

Bongard, J., Zykov, V., and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science* 314, 1118–1121. doi: 10.1126/science.1133687

Bringsjord, S., Licato, J., Govindarajulu, N. S., Ghosh, R., and Sen, A. (2015). "Real robots that pass human tests of self-consciousness," in *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Kobe), 498–504.

Cangelosi, A., and Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots.* Cambridge, MA: MIT Press.

Chalmers, D. J. (1996). *The Conscious Mind: in Search of a Fundamental Theory*. New York, NY: Oxford University Press.

Chalmers, D. J. (2008). "Foreword," in *Supersizing the Mind*, Vol. 8, ed A. Clark (Oxford: Oxford University Press), 1–33.

Chella, A., and Manzotti, R. (eds.) (2007). *Artificial Consciousness.* Exeter: Imprint Academic.

Chella, A., and Manzotti, R. (2009). Machine consciousness: a manifesto for robotics. *Int. J. Mach. Conscious.* 1, 33–51. doi: 10.1142/S1793843009000062

Chrisley, R., and Ziemke, T. (2006). *Embodiment in: Encyclopedia of Cognitive Science*. Hoboken, NJ: John Wiley and Sons, Ltd.

Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain.* New York, NY: Pantheon Books.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. London: Penguin Books.

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science* 358, 486–492. doi: 10.1126/science.aan8871

Dennett, D. C. (1978). "Where am I?" in *Brainstorms: Philosophical Essays On Mind and Psychology*, ed D. C. Dennett (Montgomery: Bradford), 317–330.

Dowe, P. (2000). *Physical Causation*. New York, NY: Cambridge University Press.

Dowe, P. (2007). Causal Processes. *Stanf. Encyclop. Philos.* Available online at: https://plato.stanford.edu/entries/causation-process/ (Accessed March 27, 2018).

Edelman, G. M., Reeke, G. N., Gall, W. E., Tononi, G., Williams, D., and Sporns, O. (1992). Synthetic neural modeling applied to a real-world artifact. *Proc. Natl. Acad. Sci. U.S.A.* 89, 7267–7271. doi: 10.1073/pnas.89.15.7267

Franklin, S. (2003). IDA - a conscious artifact? *J. Conscious. Stud.* 10, 47–66.

Franklin, S., Madl, T., D'Mello, S., and Snaider, J. (2014). LIDA: a systems-level architecture for cognition, emotion, and learning. *IEEE Trans. Auton. Ment. Dev.* 6, 19–41. doi: 10.1109/TAMD.2013.2277589

Grossberg, S. (2007). Consciousness CLEARS the mind. *Neural Netw.* 20, 1040–1053. doi: 10.1016/j.neunet.2007.09.014

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

Gunkel, D. J. (2012). *The Machine Question.* Cambridge, MA: MIT Press.

Harnad, S. (2003). Can a machine be conscious? How? *J. Conscious. Stud.* 10, 67–75.

Holland, O. (ed.). (2003). *Machine Consciousness*. New York, NY: Imprint Academic.

Holland, O. (2004). "The future of embodied artificial intelligence: machine consciousness?" in *Embodied Artificial Intelligence*, ed F. Iida (Berlin: Springer), 37–53.

Kim, J. (1989). The myth of nonreductive materialism. *Proc. Am. Philos. Soc.* 63, 31–47. doi: 10.2307/3130081

Kim, J. (1998). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Koch, C., and Tononi, G. (2008). Can machines be conscious? *IEEE Spectrum* 45, 55–59. doi: 10.1109/MSPEC.2008.4531463

Koch, C., and Tononi, G. (2017). Can we quantify machine consciousness? *IEEE Spectrum* 54, 65–69. doi: 10.1109/MSPEC.2017.7934235

Kuipers, B. (2008). Drinking from the firehose of experience. *Artif. Intell. Med.* 44, 55–70. doi: 10.1016/j.artmed.2008.07.010

Lavazza, A., and Manzotti, R. (2013). An externalist approach to creativity: discovery versus recombination. *Mind Soc.* 12, 61–72. doi: 10.1007/s11299-013-0124-6

Legg, S., and Hutter, M. (2007). "A collection of definitions of intelligence," in *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms (*Amsterdam: IOS Press), 17–24.

Manzotti, R. (2006). A process oriented view of conscious perception. *J. Consciou. Stud.* 13, 7–41.

Manzotti, R. (2017). *Consciousness and Object. A Mind-Object Identity Physicalist Theory*. Amsterdam: John Benjamins Pub.

Manzotti, R., and Chella, A. (2016). "The causal roots of integration and the unity of consciousness," in *Biophysics of Consciousness: A Foundational Approach*, eds R. R. Poznanski, J. A. Tuszynski and T. E. Feinberg (Singapore: World Scientific), 189–229.

Manzotti, R., and Tagliasco, V. (2005). From "behaviour-based" robots to "motivations-based" robots. *Rob. Auton. Syst.* 51, 175–190. doi: 10.1016/j.robot.2004.10.004

McCarthy, J. (2002). *Making Robots Conscious of Their Mental States*. Available online at: http://jmc.stanford.edu/articles/consciousness.html (Accessed March 12, 2018).

McDermott, D. (2001). *Mind and Mechanisms*. Cambridge, MA: MIT Press; Bradford Books.

Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., et al. (2012). "The iCub platform: a tool for studying intrinsically motivated learning," in *Intrinsically Motivated Learning in Natural and Artificial Systems,* eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer), 433–458.

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

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol.* 10:e1003588. doi: 10.1371/journal.pcbi.1003588

O'Regan, K., and Nöe, A. (2001). A sensorimotor account of visual perception and consciousness. *Behav. Brain Sci.* 24, 939–1011. doi: 10.1017/S0140525X01000115

Perlovsky, L. I. (2006). Toward physics of the mind: concepts, emotions, consciousness, and symbols. *Phys. Life Rev.* 3, 23–55. doi: 10.1016/j.plrev.2005.11.003

Perlovsky, L. I. (2016). Physics of the mind. *Front. Syst. Neurosci.* 10:84. doi: 10.3389/fnsys.2016.00084

Raoult, A., and Yampolskiy, R. (2015). *Reviewing Tests for Machine Consciousness*. Available online at: https://www.researchgate.net/publication/284859013_DRAFT_Reviewing_Tests_for_Machine_Consciousness (Accessed March 12, 2018).

Reeke, G. N., Sporns, O., and Edelman, G. M. (1990). Synthetic neural modeling: the "Darwin" series of recognition automata. *Proc. IEEE* 78, 1498–1530.

Reggia, J. A. (2013). The rise of machine consciousness: studying consciousness with computational models. *Neural Netw.* 44, 112–131. doi: 10.1016/j.neunet.2013.03.011

Scheutz, M. (2014). "Artificial emotions and machine consciousness," in *The Cambridge Handbook of Artificial Intelligence,* eds K. Frankish and W. Ramsey (Cambridge, MA: Cambridge University Press), 247–266.

Searle, J. R. (1984). *Minds, Brains, and Science*. Cambridge, MA: Harvard University Press.

Searle, J. R. (1990). Is the brain a digital computer? *Proc. Am. Philos. Soc.* 64, 21–37. doi: 10.2307/3130074

Seth, A. K. (2009). The strength of weak artificial consciousness. *Int. J. Mach. Conscious.* 1, 71–82. doi: 10.1142/S1793843009000086

Seth, A. K. (2010). The grand challenge of consciousness. *Front. Psychol.* 1:5. doi: 10.3389/fpsyg.2010.00005

Seth, A. K. (2016). The real problem. *Aeon*. Available online at: https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one (Accessed February 5, 2018).

Shanahan, M. (2010). *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Oxford: Oxford University Press.

Shanahan, M. P. (2005). Global access, embodiment, and the conscious subject. *J. Conscious. Stud.* 12, 46–66.

Shanahan, M. P. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Conscious. Cognit.* 15, 433–449. doi: 10.1016/j.concog.2005.11.005

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Stewart, J. (2010). "Foundational issues in enaction as a paradigm for cognitive science: from the origin of life to consciousness and writing," in *Enaction. Toward a New Paradigm for Cognitive Science,* eds J. Stewart, O. Gapenne, and E. Di Paolo (Cambridge, MA: The MIT Press), 1–31.

Stewart, J., Gapenne, O., Di Paolo, E. A., and Paolo, E. A. D. (2010). *Enaction*. Cambridge, MA: The MIT Press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42

Tononi, G. (2012). *Phi: A Voyage from the Brain to the Soul.* New York, NY: Pantheon Books.

Conflict of Interest Statement: 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.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 26 June 2018
doi: 10.3389/frobt.2018.00075

Experiments in Artificial Theory of Mind: From Safety to Story-Telling

*Alan F. T. Winfield**

Bristol Robotics Laboratory, University of the West of England, Bristol, United Kingdom

Theory of mind is the term given by philosophers and psychologists for the ability to form a predictive model of self and others. In this paper we focus on synthetic models of theory of mind. We contend firstly that such models—especially when tested experimentally—can provide useful insights into cognition, and secondly that artificial theory of mind can provide intelligent robots with powerful new capabilities, in particular social intelligence for human-robot interaction. This paper advances the hypothesis that simulation-based internal models offer a powerful and realisable, theory-driven basis for artificial theory of mind. Proposed as a computational model of the simulation theory of mind, our simulation-based internal model equips a robot with an internal model of itself and its environment, including other dynamic actors, which can test (i.e., simulate) the robot's next possible actions and hence anticipate the likely consequences of those actions both for itself and others. Although it falls far short of a full artificial theory of mind, our model does allow us to test several interesting scenarios: in some of these a robot equipped with the internal model interacts with other robots without an internal model, but acting as proxy humans; in others two robots each with a simulation-based internal model interact with each other. We outline a series of experiments which each demonstrate some aspect of artificial theory of mind.

Keywords: anticipation, simulation-based internal models, theory-of-mind, cognitive robotics, multi-robot systems, human-robot interaction, social intelligence, machine consciousness

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Ricardo Sanz,
Universidad Politécnica de Madrid (UPM), Spain
Virginia Dignum,
Delft University of Technology, Netherlands

****Correspondence:***
Alan F. T. Winfield
alan.winfield@brl.ac.uk

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *30 January 2018*
Accepted: *04 June 2018*
Published: *26 June 2018*

Citation:
Winfield AFT (2018) Experiments in Artificial Theory of Mind: From Safety to Story-Telling. Front. Robot. AI 5:75. doi: 10.3389/frobt.2018.00075

1. INTRODUCTION

Theory of mind is the term given by philosophers and psychologists for the ability to predict the actions of self and others (Carruthers and Smith, 1996). With theory of mind, it is supposed, we are able to anticipate how others might behave in particular circumstances. However, the idea of theory of mind is empirically weak—we have only a poor understanding of the neurological or cognitive processes that give rise to theory of mind. Artificial Intelligence (AI), and its embodied counterpart—robotics, provides a powerful synthetic approach to theory of mind because it allows us to ask the question "how would we build artificial theory of mind in a robot?" and opens the possibility that we could test theories of theory of mind.

The role of theory of mind in consciousness (or, indeed of consciousness in theory of mind) is both unclear and controversial (Carruthers, 2009; Sebastian, 2016). In this paper we avoid this difficult question by focusing instead on synthetic models of theory of mind. We contend firstly that such models—especially when tested experimentally—can provide valuable insights into both natural and artificial cognition, and secondly that artificial theory of mind can provide intelligent robots with powerful new capabilities, in particular social intelligence for human-robot interaction. Artificial theory of mind has been recently highlighted as one of the *Grand Challenges of Science Robotics*: "The three most significant challenges that stem from building robots that interact socially with people are modeling social dynamics, learning social and moral norms, and building a robotic theory of mind" (Yang et al., 2018).

The aim of this paper is to advance the hypothesis that simulation-based internal models offer a powerful and realizable, theory-driven basis for artificial theory of mind. Proposed as a computational model of the simulation theory of mind (Goldman, 2006), our simulation-based internal model equips a robot with an internal model of itself and its environment, including other dynamic actors, which can test (i.e., simulate) the robot's next possible actions and hence anticipate the likely consequences of those actions both for itself and others; importantly our simulation-based internal model is a practical proposition with current technology. Although it falls far short of a full artificial theory of mind, our model allows us to test several interesting scenarios: in some of these a robot equipped with the internal model interacts with other robots, without an internal model but acting as proxy humans; in others two robots each with a simulation-based internal model interact with each other. We are able to predict second and third order interactions[1] and, in some cases, observe interesting and unexpected emergent behaviors.

This paper proceeds as follows. First in section 2 we adopt a working definition of theory of mind and outline theories of theory of mind. Choosing the simulation theory of mind (ST) we then outline the conceptual basis for simulation-based internal models, together with prior work which uses such models, before proposing a generic computational model of ST. Section 3 then introduces a series of experiments in (simple) artificial theory of mind: in the first the aim is improved *safety*; in the second it is simple *ethical* behaviors—including a scenario in which the ethical robot faces a dilemma; in the third one robot aims to infer the goals of another to rationally *imitate* it. The fourth and final exemplar is a thought experiment which outlines a proposal for an embodied computational model of *storytelling*, using robots. It is important to note that none of these experiments were conceived as a solution to the problem of artificial theory of mind. It was only *post-hoc* that we recognized that—since each experiment involves one or more robots which predict the behavior of others—taken together they offer some insight into practical artificial theory of mind. Take ethical robots as a case in point. Although a robot may not need theory of mind to behave ethically it is easy to see that the ability to predict the intentions of others would greatly facilitate and likely extend the scope of its ethical responses[2]. Section 4 concludes the paper with a discussion which both draws high-level conclusions from the experimental work outlined in section 3 and makes the case that this work does demonstrate a number of components of theory of mind and can therefore reasonably be described as "experiments in artificial theory of mind."

2. FROM SIMULATION THEORY TO A SIMULATION-BASED INTERNAL MODEL

2.1. Theories of Theory of Mind

One difficulty of this paper is that there is no single definition of theory of mind and its attributes. Definitions vary according to the context so, in animal cognition, for instance, Roberts (2001) writes "The term theory of mind refers to the fact that people know about minds ... the inferences you make about others minds may often guide your behavior," whereas Breed and Moore (2012) write "An animal with a theory of mind can form hypotheses about the thoughts of surrounding animals." In child development theory of mind refers to "childrens understanding of people as mental beings, who have beliefs, desires, emotions, and intentions" (Astington and Dack, 2008), with mental representation and false belief regarded as key components. And in Birch et al. (2017) "Perspective taking, or theory of mind, involves reasoning about the mental states of others (e.g., their intentions, desires, knowledge, beliefs) and is called upon in virtually every aspect of human interaction." In this paper we resolve this difficulty by settling on "to explain and predict the actions, both of oneself, and of other intelligent agents" as our working definition of theory of mind (Carruthers and Smith, 1996).

There are a number of theories of theory of mind (Carruthers and Smith, 1996); and such theories are generally grouped into two broad categories, known as theory theory (TT) and simulation theory (ST)[3]. For a good outline comparison of TT and ST see Michlmayr (2002). Theory theories hold that one intelligent agent's understanding of another's mind is based on innate or learned rules, sometimes known as folk psychology. In TT these hidden rules constitute a "theory" because they can be used to both explain and make predictions about others' intentions. In contrast "simulation theory suggests that we do not understand others through the use of a folk psychological theory. Rather, we use our own mental apparatus to form predictions and explanations of someone by putting ourselves in the shoes of another person and simulating them" (Michlmayr, 2002). Goldman (2006) introduces the idea of mental simulation: "the simulation of one mental process by another mental process," and makes the important distinction between intra personal and interpersonal mental simulation; the former is simulation of self, and the latter the simulation of other. Goldman (2006) also marks the distinction between computational modeling simulation and replication simulation, noting that only the latter is of interest to theory of mind; we would contend that the former is of great interest to *artificial* theory of mind.

In this paper we adopt ST as both the inspiration and theoretical basis for our hypothesis that simulation-based internal models offer a powerful approach to building artificial theory of mind, not because we have a principled theoretical preference for ST over TT, but because simulation-based internal models provide a realizable computational model for ST.

Also relevant here is the simulation theory of cognition (Hesslow, 2002; Wilson, 2002). This theory hypothesizes that cognitive introspection utilizes the same processes as interaction with the external environment. During introspection (thinking), actions are covert and are assumed to generate, via associative brain mechanisms, the sensory inputs that elicit further actions (Hesslow, 2012). In this view, cognition requires a grounded

[1] Second order interactions are between robot and environment and third order are robot-robot interactions.

[2] While noting that having a theory of mind does not make an agent ethical

[3] There are also a number of hybrid theories which combine elements of TT and ST.

representation of the world that is not composed of abstract symbols; a simulation provides just such a model.

2.2. Simulation-Based Internal Modeling

A simulation-based internal model is a mechanism for internally representing both the system and its current environment. If we embed a simulation of a robot, including its currently perceived environment, inside that robot then the robot has a "mechanism for generating and testing *what-if* hypotheses; i.e.,

1. *what if* I carry out action *x*..? and, ...
2. of several possible next actions x_i, *which* should I choose?" (Winfield, 2014)

Holland writes: an Internal Model allows a system to look ahead to the future consequences of current actions, without actually committing itself to those actions (Holland, 1992, p. 25). This leads to the idea of "an internal model as a *consequence engine*—a mechanism for predicting and hence anticipating the consequences of actions" (Winfield and Hafner, 2018).

The idea of embedding a simulator of a robot within that robot is not new, but implementation is technically challenging, and there have been relatively few examples described in the literature. One notable example is within the emerging field of machine consciousness (Holland, 2003; Holland and Goodman, 2003). Marques and Holland (2009) define a "functional imagination" as "a mechanism that allows an embodied agent to simulate its own actions and their sensory consequences internally, and to extract behavioral benefits from doing so"; a embedded simulation-based internal model provides such a mechanism.

Bongard et al. (2006) describe a 4-legged starfish like robot that makes use of explicit internal simulation, both to enable the robot to learn it's own body morphology and control, and notably allow the robot to recover from physical damage by learning the new morphology following the damage. The internal model of Bongard et al. models only the robot, not its environment. See also Zagal and Lipson (2009). In contrast Vaughan and Zuluaga (2006) demonstrate self-simulation of both a robot and its environment in order to allow a robot to plan navigation tasks with incomplete self-knowledge; their approach significantly provides perhaps the first experimental proof-of-concept of a robot using self-modeling to anticipate and hence avoid unsafe actions. Zagal et al. (2009) describe self-modeling using internal simulation in humanoid soccer robots; in what they call a 'back-to-reality' algorithm, behaviors adapted and tested in simulation are transferred to the real robot.

In robotics advanced physics and sensor-based simulation tools are routinely used to model, develop or evolve robot control algorithms prior to real-robot tests. Well-known robot simulators include Webots (Michel, 2004), Gazebo (Koenig and Howard, 2004), Player-Stage (Vaughan and Gerkey, 2007), and V-REP (Rohmer et al., 2013). Simulation technology is now sufficiently mature to provide a practical route to implementation of an embedded simulation-based internal model. Furthermore Stepney (2018) sets out a principled approach to simulation which treats a simulator as a scientific instrument.

2.3. A Computational Model of Simulation Theory of Mind

We have recently proposed an architecture for a robot with a simulation-based internal model which is used to test and evaluate the consequences of that robot's next possible actions. Shown in **Figure 1** "the machinery for modeling next actions is relatively independent of the robot's controller; the robot is capable of working normally without that machinery, albeit without the ability to generate and test what-if hypotheses. The what-if processes are not in the robot's main control loop, but instead run in parallel to moderate the Robot Controller's normal action selection process, acting in effect as a kind of governor" (Blum et al., 2018). This governance might be to rule out certain actions because they are modeled as unsafe for the robot, or to recommend new robot actions to, for instance, prevent an accident.

"At the heart of the architecture is the Consequence Engine. The CE is initialized from the Object Tracker-Localizer, and loops through all possible next actions; these next actions are generated within the Robot Controller (RC) and transferred to the mirror RC within the CE (for clarity this data flow is omitted from **Figure 1**). For each candidate action the CE simulates the robot executing that action, and generates a set of model outputs ready for evaluation by the Action Evaluator. The Consequence Evaluator loops through each possible next action; this is the Generate-and-Test loop. Only when the complete set of next possible actions has been tested does the Consequence Evaluator send, to the Robot Controller, its recommendations" (Winfield et al., 2014). These processes are explained in detail in Blum et al. (2018).

We argue that the architecture outlined here represents a computation model of artificial theory of mind. First, the model clearly provides a robot with the ability to self-model and hence predict the consequences of its own actions. Second the model can be used to predict another dynamic agent's actions and—if they interact—the consequences of this robot's actions to that other agent. This predictive modeling of others can be implemented in two ways depending on the way we model those other agents.

1. In the first, which we can call the ST-self plus TT-other (ST+TT) model, the other dynamic agents (i.e., robots) are modeled within the World Model of this robot using simple theory, for example a ballistic model for moving agents. Since this variant combines elements of ST and TT it models a hybrid theory of mind.
2. In the second, which we can call ST-self plus ST-other (ST+ST), the whole of the consequence engine can be initialized for the other agent and run introspectively, recalling the simulation theory of cognition (Hesslow, 2012). Here the robot models each other agent *exactly* as it models itself, i.e., as a conspecific. This variant models pure ST[4].

The experiments outlined in the next section illustrate both ST+TT and ST+ST variants.

[4] Noting that even our computational model of ST is not completely theory free, since the world model models the physics of collisions, etc.

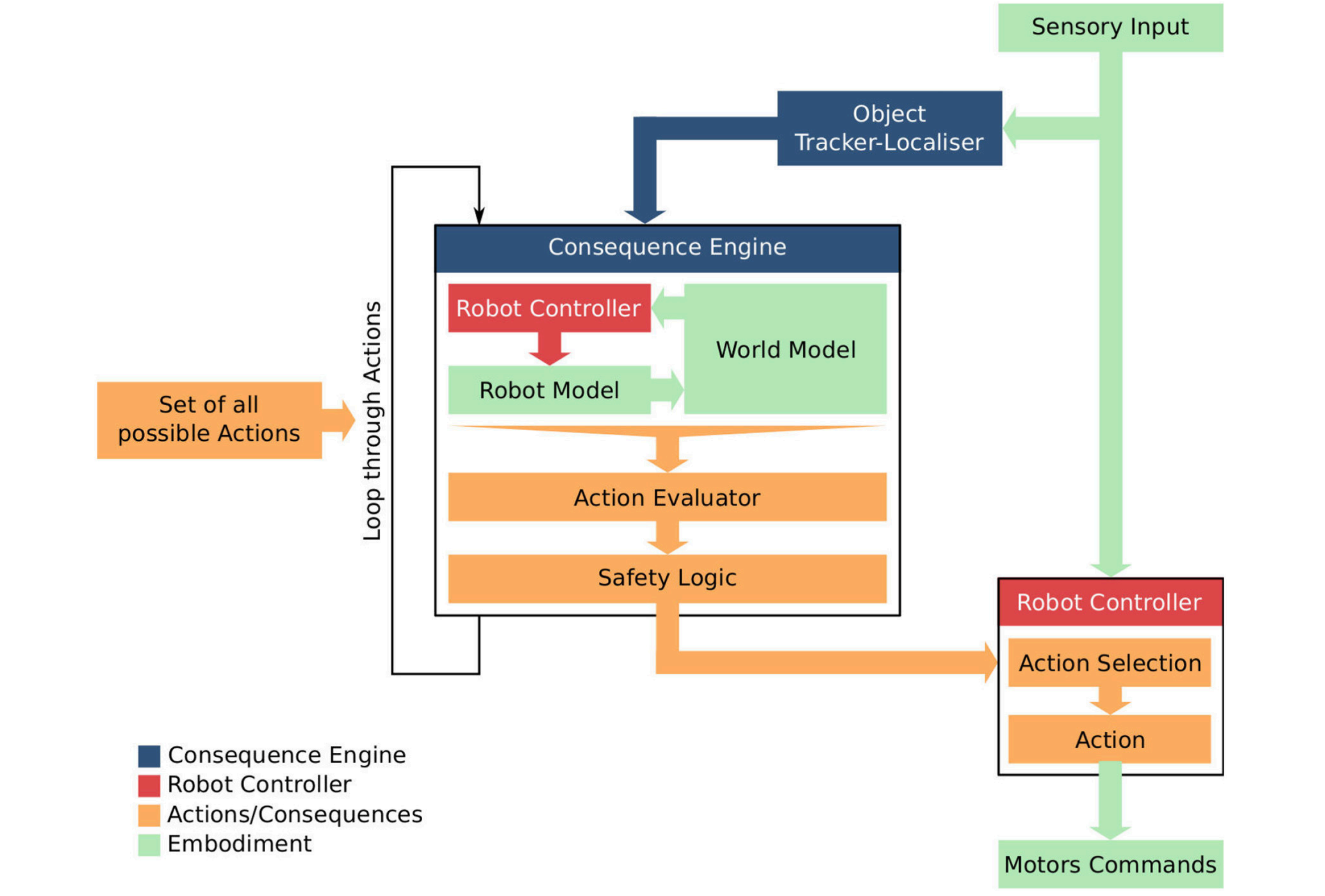


FIGURE 1 | The Consequence Engine: an architecture for robot anticipation using a simulation-based internal model. Figure from Blum et al. (2018).

3. EXPERIMENTS IN ARTIFICIAL THEORY OF MIND

3.1. Safety: The Corridor Experiment

We have implemented and tested the simulation-based internal model architecture outlined above in an experimental scenario, which we call the corridor experiment (Blum et al., 2018). Inspired by the problem of how mobile robots could move quickly and safely through crowds of moving humans, the aim of this experiment is to compare the performance of our simulation-based internal model with a purely reactive approach. In other words: can a robot's safety be improved with simple artificial theory of mind?

In this experiment one mobile robot (the CE-robot) is equipped with the consequence engine of **Figure 1**, while 5 other mobile robots have only simple obstacle avoidance behaviors. The setup is shown in **Figure 2** (left); here the smart CE-robot is shown in blue at its starting position. The CE-robot's goal is to reach the end of the corridor on the right while maintaining its own safety by avoiding—while also maintaining a safe distance—the five proxy-human robots shown in red. **Figure 2** (right) shows the trajectories of all six robots during a simulated run of the experiment, with the CE-robot reaching the end of the corridor. **Figure 3** shows the real-robot experimental setup.

In this experiment the CE robot models each of the proxy-human robots as a ballistic agent with obstacle avoidance—in other words as agents that will continue to move in their current direction and speed unless confronted with an obstacle, which may be another agent or the corridor wall. The CE runs in real-time and is updated every 0.5 s with the actual position and direction of the proxy-humans within the CE robot's attention radius. This is not an unreasonable model when considering how you might behave when avoiding another person who is not paying attention to where they are going—peering at their smartphone perhaps.

Results of the corridor experiment (detailed in Blum et al., 2018) show that for a relatively small cost in additional distance covered, the likelihood that a proxy-human robot comes within the CE-robot's safety radius falls to zero. Clearly there is a computational cost. This is entirely to be expected: anticipatory modeling of other agents clearly incurs a computational overhead.

In the corridor experiment there is an asymmetry: the CE-robot has a model for the proxy-human robots whereas they

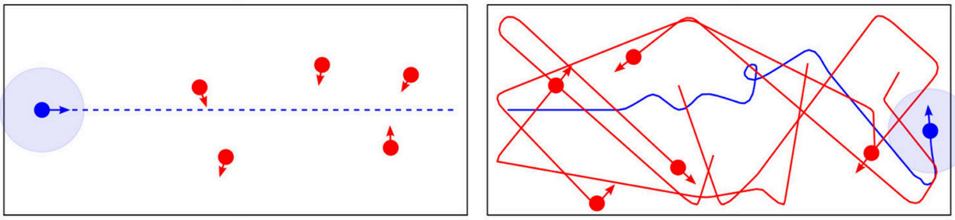


FIGURE 2 | The corridor experiment goal **(left)**, with 5 (red) robots moving randomly and one intelligent (CE) robot (blue) with a simulation-based internal model. **(Right)** shows (simulated) trajectories of all six robots by the time blue has reaching the end of the corridor. Figure from Blum et al. (2018).

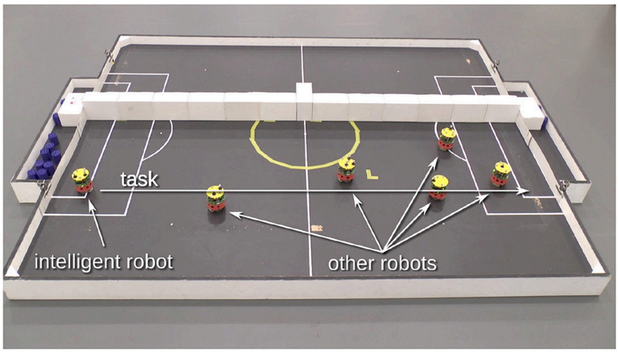


FIGURE 3 | The corridor experiment, using e-puck robots (Mondada et al., 2009) fitted with Linux extension boards (Liu and Winfield, 2011). This image shows the initial condition with the CE (intelligent) robot on the left and the five proxy-human robots positioned at randomly selected locations in the corridor. The arena markings have no significance here. Figure from Blum et al. (2018).

have no model for the CE-robot. In an extension to the corridor experiment which we call the pedestrian experiment two robots—each equipped with the same CE—approach each other. As with the corridor experiment each models the other as a simple ballistic agent but here we have symmetry with each agent paying full attention to the other, trying to anticipate how it might behave and planning its own actions accordingly. Is it possible that our "pedestrian" robots might, from time to time, engage in the kind of "dance" that human pedestrians do when one steps to their left and the other to their right only to compound the problem of avoiding a collision with a stranger?

Results show that we do indeed observe this interesting emergent behavior. In five experimental runs four resulted in the two pedestrian robots passing each other by both turning either to the left or to the right—**Figure 4** (left) shows one example of this behavior. However, in one run, shown in **Figure 4** (right) we observe a brief dance caused when both robots decide, at the same time, to turn toward each other—each predicting wrongly that the other robot would continue its currently trajectory—before the two robots resolve the impasse and pass each other safely.

3.2. Toward Ethical Robots

We have conducted exploratory work—based on the same simulation-based internal model architecture outlined in section 2—to explore the possibility of robots capable of making decisions based on ethical rules. These robots implement simple consequentialist ethics with rules based on Asimov's famous laws of robotics. Following Asimov's first law: "a robot may not harm a human or, through inaction, allow a human to come to harm," our ethical robot will act proactively when it anticipates (a) that a proxy-human robot is in danger of coming to harm and (b) the ethical robot can itself intervene. We have experimentally tested such a minimally ethical robot initially with e-puck robots (Winfield et al., 2014) and subsequently with NAO humanoid robots (Vanderelst and Winfield, 2018). As in the corridor experiment the ethical robot's CE models the proxy-human(s) as simple ballistic agents. In some experiments we have extended those TT models so that the ethical robot can, for instance, call out "danger!" and if the human robot then responds with "ok, understood" the ethical robot will change its model for that human from "irresponsible" to "responsible" and not intervene as it heads toward the danger zone. In this way the ethical robot is able to modify its belief about the proxy-human.

Figure 5 shows results from one trial with two NAO humanoid robots, one (blue) equipped with a CE and ethical logic layer, and the other (red) programmed only with short range obstacle avoidance behavior to act as a proxy-human. **Figure 5** shows that the ethical robot does indeed reliably intervene, diverting from its own path, and when red halts to avoid a collision with blue, blue then continues toward its own goal.

We have tested the same ethical robot (running identical code) in a scenario with two proxy-humans both heading toward danger at the same time. These trials, first with e-puck robots (Winfield et al., 2014) and more recently with NAO robots, are believed to be the first experimental tests of a robot facing an ethical dilemma. We did not provide the ethical robot with a rule or heuristic for choosing which proxy-human to "rescue" first, so that the ethical robot faces a balanced dilemma. **Figure 6** (left) shows the experimental arena with the ethical robot (blue) initially equidistant from the two (red) proxy-human robots. The trajectory plots in **Figure 6** (right) interestingly show that in three of the five trials blue initially chose to move toward the red robot heading toward danger (B), but then appeared to 'change its mind' to "rescue" the other red robot. Exactly the same "dithering" emergent behavior was observed with the e-puck robots in Winfield et al. (2014), and can be explained in part by the fact that the ethical robot's CE is running continuously, re-evaluating the consequences of its own and the other robots'

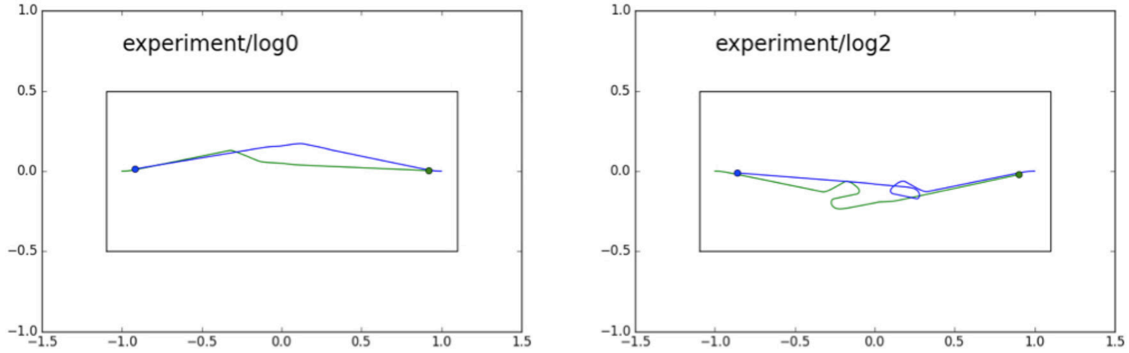


FIGURE 4 | The pedestrian experiment—two trials showing robot trajectories. Two robots, blue and green, are each equipped with a CE. Blue starts from the right, with a goal position on the left, while at the same time green starts from the left with a goal position on the right. **(Left)** We see the typical behavior in which the two robots pass each other without difficulty, normally because one robot—anticipating a collision—changes direction first, in this case green. **(Right)** Here both robots make a decision to turn at the same time, green to its left and blue to its right; a "dance" then ensues before the impasse is resolved.

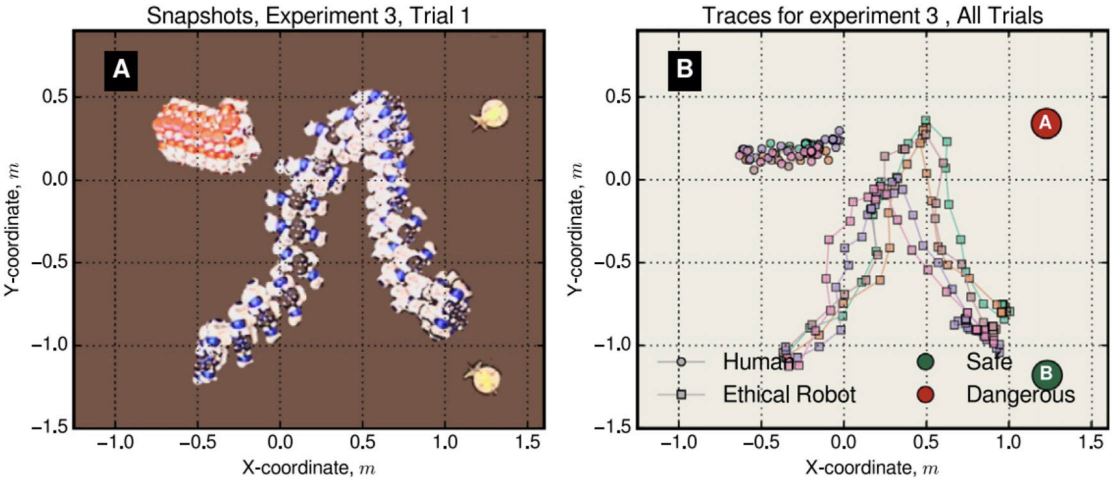


FIGURE 5 | An ethical humanoid robot (blue) anticipates that proxy-human robot (red) is heading toward danger (location A at the top right). It diverts from its path toward goal position B (bottom right) to intersect red's path. Red then stops and blue resumes its path toward its goal. **(A)** Shows the trajectories of Blue and Red for trial 1. **(B)** Shows all 5 experimental trials. Figure from Vanderelst and Winfield (2018).

behaviors and perhaps choosing a new action once per second[5]. This makes our ethical robot pathologically indecisive.

3.3. The Imitation of Goals

The imitation of goals is a very important form of social learning in humans. This importance is reflected in the early emergence of imitation in human infants; from the age of two, humans can imitate both actions and their intended goals (Gariépy et al., 2014) and this has been termed rational imitation.

Imitation has long been regarded as a compelling method for (social) learning in robots. However, robot imitation faces a number of challenges (Breazeal and Scassellati, 2002). One of the most fundamental issues is determining what to imitate (Carpenter et al., 2005). Although not trivial it is relatively straightforward to imitate actions, but inferring goals from observed actions and thus determining which parts of a demonstrated sequence of actions are relevant, i.e., rational imitation, is a difficult research problem.

The approach we explore in Vanderelst and Winfield (2017), is to equip the imitating robot with a simulation-based internal model that allows the robot to explore alternative sequences of actions required to attain the demonstrator robot's potential goals (i.e., goals that are possible explanations for the observed actions). Comparing these actions with those observed in the demonstrator robot enables the imitating robot to infer the goals underlying the observed actions.

Figure 7 shows one of several experiments from Vanderelst and Winfield (2017). Here the red robot imitates the goals of the blue robot. In condition 1 blue moves directly to its goal position

[5] In practice the ethical NAO robot would also favor the slower of the two proxy-humans, as a side-effect of its action-selection logic.

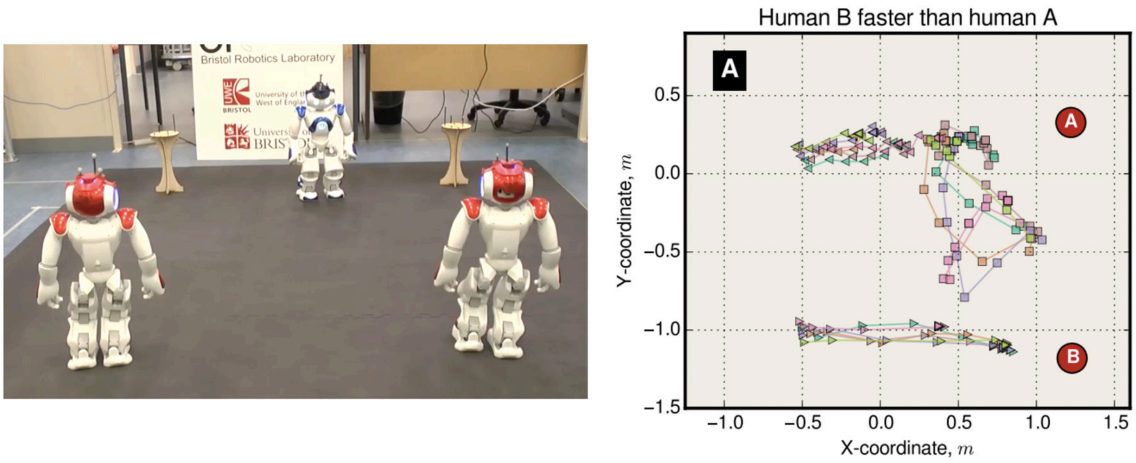


FIGURE 6 | An ethical dilemma. **Left**: The ethical robot is initially positioned midway between and slightly to the front of two danger zones A and B. **Right**: The ethical robot's trajectories are shown here plotted with squares. Two proxy-human robots start from the left, both heading toward danger—trajectories plotted with triangles. Results of 5 trials are shown here.

(**Figures 7A,B**). Blue infers the goal is to move to red's goal and does so directly in **Figure 7C**. In condition 2 blue deviates around an obstacle even though it has a direct path to its goal (**Figures 7D,E**). In this case red infers that the deviation must be a sub-goal of blue—since blue is able to go directly to its goal but chooses not to—so in **Figure 7F** red creates a trajectory via blue's sub-goal. In other words red has inferred blue's intentions to imitate its goals. In condition 3 blue's path to its goal is blocked so it has no choice but to divert (**Figures 7G,H**). In this case red infers that blue has no sub-goals and moves directly to the goal position (**Figure 7I**).

3.4. An Embodied Computational Model of Storytelling

Consider the idea that some of the what-if sequences tested with a robot's consequence engine are constructed fictions, i.e., "if I had turned left I would have collided with a wall." While others—the ones actually enacted—could be historical narratives, i.e., "I turned right and reached my goal."

Assume that we have two robots, each equipped with the same simulation-based internal model of **Figure 1**. Let us also assume that the robots are of a similar type, in other words they are conspecifics. Let us now extend the robots' capabilities in the following way. Instead of simply discarding ("forgetting") an action that has been modeled, the robot may transmit that action and its predicted or actual consequences to another robot.

Figure 8 illustrates robot A "imagining" a what-if sequence, then narrativizing that sequence. It literally signals that sequence using some transmission medium. Since we are building a model and it would be very convenient if it is easy for human observers to interpret the model, let us code the what-if sequence verbally and transmit it as a spoken language sequence. Technically this would be straightforward to arrange since we would use a standard speech synthesis process. Although it is a trivial narrative robot A is now able to both "imagine" and then literally tell a story. If that story is of something that has not happened it is a fictional narrative, otherwise it is a historical narrative.

Robot B is equipped with a microphone and speech recognition process it is thus able to "listen" to robot As story, as shown in **Figure 9**. Let us assume it is programmed, so that a word used by A signifies the same part of the what-if action sequence to both A and B. Providing the story has been heard correctly then robot B will interpret robot A's story as a what-if sequence. Now, because robot B has the same internal modeling machinery as A- they are conspecifics- it is capable of running the story it has just heard within its own internal model. In order that this can happen we need to modify the robots programming so that the what-if sequence it has heard and interpreted is substituted for an internally generated what-if sequence. This would be easy to do. But, once that substitution is made, robot B is able to run A's what-if sequence (its story) in *exactly* the same way it runs its own internally generated next possible actions, simulating and evaluating the consequences. Robot B is therefore able to "imagine robot A's story[6].

In this model we have, in effect, co-opted the cognitive machinery for testing possible next actions for "imagining," or introspectively experiencing, heard stories. By adding the machinery for signaling and signifying internally generated sequences (narratives)—the machinery of semiotics—we have constructed an embodied computational model of storytelling.

A major problem with human-robot interaction is the serious asymmetry of theory of mind (Winfield, 2010). Consider an elderly person and her care robot. It is likely that a reasonably sophisticated near-future care robot will have a built-in (TT) model of an elderly human (or even of a particular human). This places the robot at an advantage because the elderly person has no theory of mind at all for the robot, whereas the robot has a (likely limited) theory of mind for her. Actually the situation

[6] Where is the meaning? It could be argued that when the listener replays the story in its internal model (functional imagination) that *is* meaning.

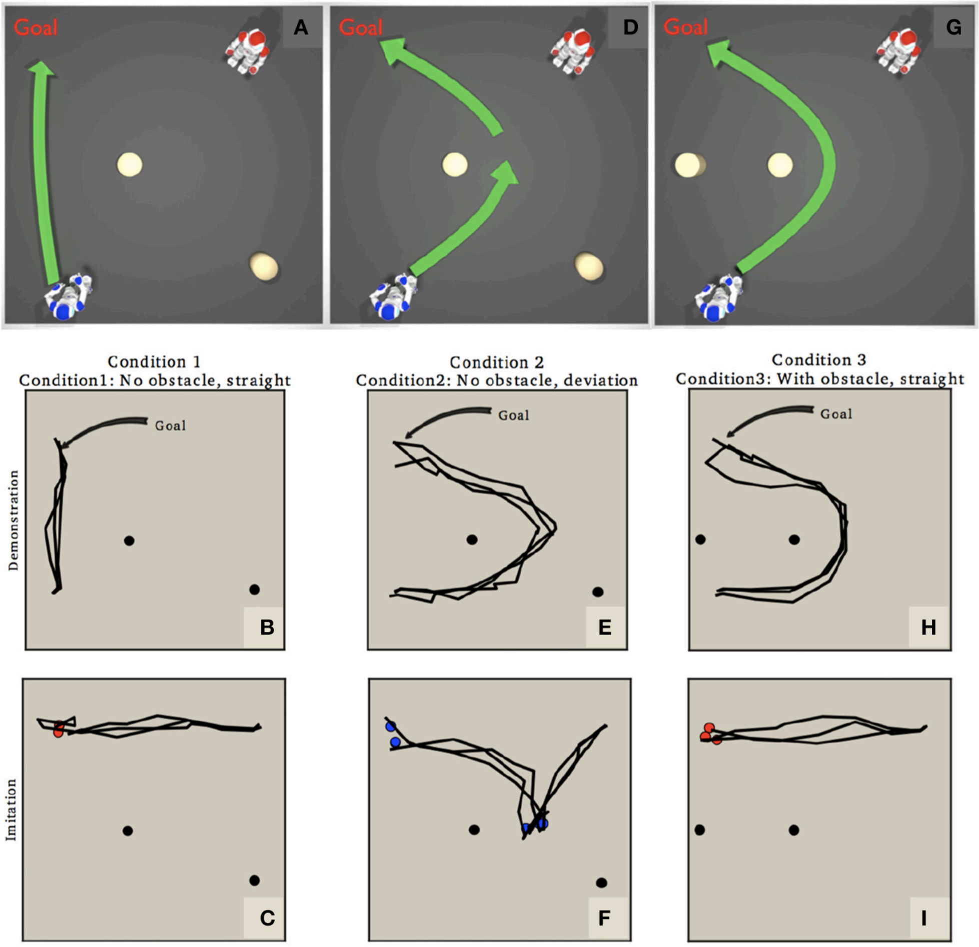


FIGURE 7 | Rational imitation. **(A,D,G)** Show the setup with blue as the demonstrating robot and red the observing (then imitating) robot. In condition 1 **(A,B,C)** blue moves directly to its goal position. In condition 2 **(D–F)** blue diverts around an obstacle even though it could move directly to its goal position. And in condition 3 **(G–I)** blue's path is blocked so it cannot go directly to its goal. **(B,E,H)** Show trajectories of 3 runs of the demonstrator robot blue, and **(C,F,I)** Show trajectories of 3 runs of the imitating robot red. Note that red starts from the position it observes from. Figures from Vanderelst and Winfield (2017).

may be worse than this, since our elderly person may have a completely incorrect theory of mind for the robot, perhaps based on preconceptions or misunderstandings of how the robot should behave and why. Thus, when the robot actually behaves in a way that doesn't make sense to the elderly person, her trust in the robot will be damaged and its effectiveness diminished (Stafford et al., 2014).

The storytelling model proposed here provides us with a powerful mechanism for the robot to be able to generate *explanations* for its actual or possible actions. Especially important is that the robot's user should be able to ask (or press a button to ask) the robot to explain "why did you just do that?" Or, pre-emptively, to ask the robot questions such as "what would you do if I fell down?" Assuming that the care robot is equipped with an autobiographical memory[7], the first of these questions would require it to re-run and narrate the most recent action sequence to be able to explain why it acted as it did, i.e., "I turned left because I didn't want to bump into you." The second kind of pre-emptive query requires the robot to interpret the question in such a way it can first initialize its internal model to match the situation described, run that model, then narrate the actions it predicts it would take in that situation. In this case the robot acts first as the listener in **Figure 9**, then as the narrator in **Figure 8**. In this way the robot would actively assist its human user to build a theory-of-mind for the robot.

4. DISCUSSION

4.1. Related Work

One of the most influential works to date on proposing and implementing artificial theory of mind is Scassellati's 2002 paper *Theory of Mind for a Humanoid Robot* (Scassellati, 2002). Based

[7] It would be relatively easy for a robot to build a memory of everything that has happened to it, but of much greater interest here is to integrate the autobiographical memory into the internal model, perhaps leading to what Conway (2005) describes as a self-memory system (SMS).

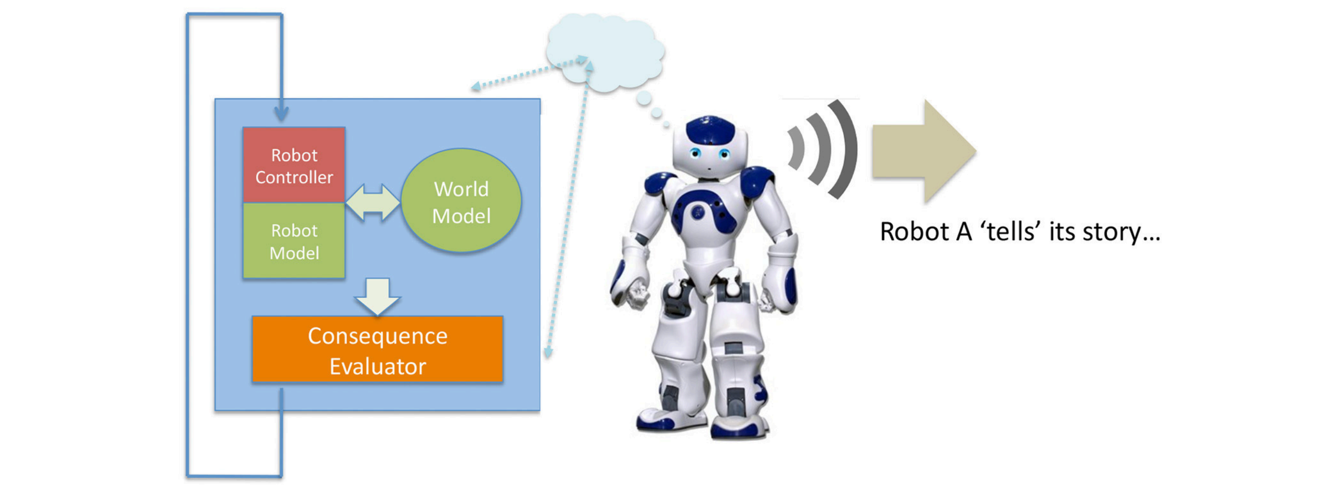


FIGURE 8 | Robot A, the storyteller, "narrativizes" one of the "what-if" sequences generated by its generate-and-test machinery. First an action is tested in the robot's internal model (left), second, that action—which may or may not be executed for real—is converted into speech and spoken by the robot. From Winfield (2018).

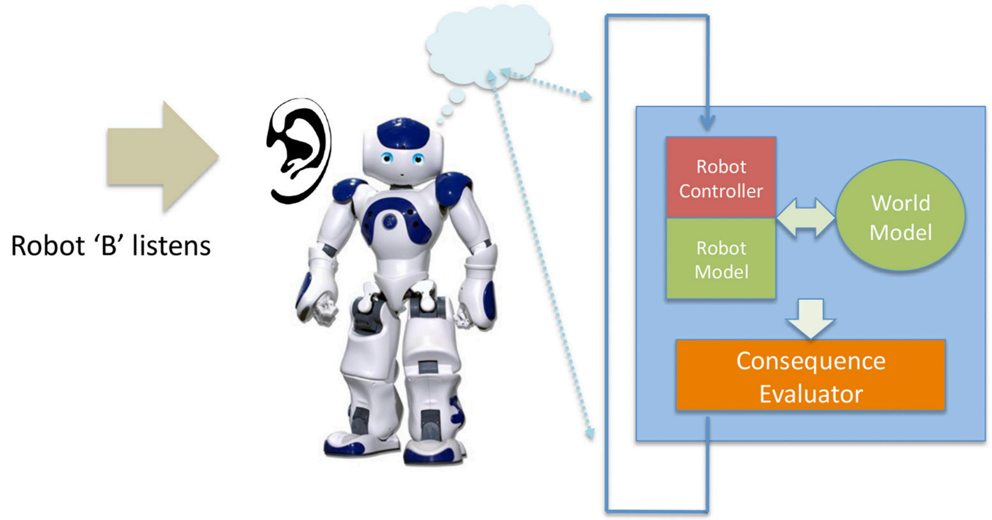


FIGURE 9 | Robot B, the listener, uses the same "what-if" cognitive machinery to "imagine" robot A's story. Here the robot hears A's spoken sequence, then converts it into an action which is tested in B's internal model. From Winfield (2018).

on aspects of theory of mind present in young (4 month old) infant humans the author describes an implementation of visual attention, finding faces and the recognition and tracking of eyes, and discrimination between animate and inanimate, on the MIT Cog robot (Brooks et al., 1999). In contrast with the present work Scassellati (2002) is based on theory theories of mind (TT). Other works have also explored the important role of shared attention in social interaction and development, for instance Deák et al. (2001) and Kaplan and Hafner (2006).

Kim and Lipson (2009) describe an approach in which one robot uses an ANN to learn another's intentions based on its behavior. A very recent paper *Machine Theory of Mind* also describes a machine learning approach in which one agent observes another's behaviors and learns a predictive model of that agent (Rabinowitz et al., 2018); the simulated agents of this work learn the rules underlying the behavior of the observed agent, hence this is also a TT approach.

Several authors have proposed artificial theory of mind as a mechanism for improved human-robot interaction. Devin and Alami (2016), for instance, describe an implementation in which a robot estimates the status of the goals of a human with which it is interacting (i.e., "in progress," "done," "aborted" or "unknown"). Görür et al. (2017) also propose a mechanism for estimating a human's beliefs about possible actions in a shared human-robot task; they propose a stochastic approach in which a Hidden Markov Model estimates action states in the set ("not ready," "ready," "in progress," "help needed," "done," and "aborted").

A number of authors focus on the role of deception as an indicator of theory of mind. Terada and Ito (2010) outline an experiment to deceive a human about the intentions of a robot, noting that the experimental result indicated that unexpected change of a robot behavior gave rise to an impression of being deceived by the robot. Wagner and Arkin (2011) describe an experiment in which two robots play a game of hide and seek in which one, the hider, attempts to deceive the seeker by sending false information.

A small number of works have also proposed "like-me" or "self-as" simulation approaches, including Kennedy et al. (2009) and Gray and Breazeal (2014). Kennedy et al. (2009) promote like-me simulation as "a powerful mechanism because for any "individual" strategy the agent has, it can reason about another agent having that strategy and, further, by creating hypothetical situations ... it can predict the actions it would take under hypothetical conditions;" the paper describes a like-me simulation based on the ACT-R/E (Adaptive Control of Thought-Rational/Embodied) architecture, with two robots in which one acts as a proxy human. Gray and Breazeal (2014) describe a very elegant experiment in which a robot simulates both its own possible actions and a human's likely perception of those actions in order to choose actions that manipulate the human's beliefs about what the robot is doing — and thereby deceive the human. These two works model the simulation theory of mind (ST) and are therefore of particular relevance to the present paper.

4.2. Discussion and Conclusions

To what extent do any of the experiments outlined in this paper demonstrate (artificial) theory of mind, as variously defined in section 2.1? We can certainly be clear about which aspects of theory of mind we cannot emulate. Our robots do not "know about minds" (Roberts, 2001) (arguably they do not know about anything), but we would also suspect that while animals have minds they too do not know about them. Nor do our robots either have, or model, affective states. And we can be quite sure that none of the robots described in this paper would pass Premack and Woodruff (1978)'s famous tests which controversially demonstrated that chimpanzee have theory of mind.

Many accounts of theory of mind are couched in terms of modeling or predicting the "mental states" of others (Astington and Dack, 2008; Birch et al., 2017), but there are two problems with the use of this term. The first is that there is no clear understanding or agreement over what mental states are in animals and humans; it seems that the term is used as a proxy for several things including beliefs, desires, emotions and intentions. Secondly, robots are not generality regarded as having mental states. They certainly do not have emotions, but they arguably can have a machine analog of simple beliefs (i.e., that the path to the left is safe, whereas the path to the right is unsafe, or a belief that another agent is moving toward danger and that by inference its mental state is "unaware of danger"), simple desires (i.e., to maintain its energy level by returning to a recharging station whenever its battery charge drops below a certain level) and intentions (i.e., goals, such as "navigate safely to position x"). Although we have not used the term mental states in this paper nor do the experiments of this paper explicitly label such states they can be properly described as predicting and/or inferring the beliefs, desires and intentions of both themselves and others.

If we accept simulation of self and other as an artificial analog of mental representation, then our robots do demonstrate this attribute. The experiments of sections 3.1 and 3.2 show that a robot with a simulation-based internal model is capable of predicting the consequences of its actions for both itself and one or more robots acting as proxy humans, and choosing actions on the basis of either safety or ethical considerations. They can therefore "reason about," i.e., model, the intentions of others, even though those models are very simple ballistic TT models and, in the case of the ethical robot experiments in section 3.2, also modeled by default as irresponsibly unaware of danger. Of course our robots have a much better model of themselves than others—but is that not also true of human theory of mind? For sure we have detailed models for those close to us—family and close friends—but our models of strangers, when walking on a sidewalk for example, can be very simple (Helbing and Molnar, 1995).

Although it is an unsophisticated example, arguably the pedestrian experiment in section 3.1 demonstrates false beliefs when each models the other as continuing in a straight line when, in fact, they each turn into the other's path (**Figure 4**, right). In fact we have also shown that it is surprisingly easy to turn an ethical robot into a mendacious (deceptive) robot, so that it behaves either competitively or aggressively toward a proxy human robot (Vanderelst and Winfield, 2016).

We have also demonstrated, in section 3.3, that a robot with a simulation-based internal model can infer the goals of another robot, therefore learning the other robot's intentions. Imitation is a powerful form of social learning and we argue that the inferential learning of section 3.3 demonstrates another key component of theory of mind.

The model of storytelling proposed in section 3.4 gets, we contend, to the heart of theory of mind. Theory of mind works best between conspecifics: in general you can much better understand your partner's beliefs and intentions than your cat's. The two robots in our thought experiment of section 3.4 would in principle be able to learn each other's beliefs and intentions in a very natural (to humans) way, through *explanation*. This is, after all, one of the key mechanisms by which infant humans learn theory of mind; one only has to think of a child asking "Mummy why are you angry with me?" (Ruffman et al., 2002).

The robots of this paper all have the cognitive machinery to predict their own behavior. But we must not assume that because a robot can predict its own behavior it can predict the behavior of any other agent. Of course when those others are conspecifics then predicting the behavior of others 'like me' becomes a (conceptually) straightforward matter of co-opting your own internal model to model others'. In all of the experiments of this paper we make use of homogeneous robots, which clearly share the same architecture (although in some cases those robots are programmed to behave differently, as proxy humans for instance). In a heterogeneous multi-robot system a robot might need to model the beliefs or intentions of a robot quite unlike itself, and the same is clearly true for a robot that might need to model the mental states of a human. But as Gray and Breazeal

TABLE 1 | Table summarizing the contribution of each of the experiments of section 3 together with their respective theory modes (as defined in section 2.3).

Experiment	Figures	Theory mode (section 2.3)	Notes
Corridor experiment	2	ST+TT	One robot with ST model of self and ballistic TT model of five other robots, demonstrates predictive modeling of self and reasoning about the intentions of others, and attention radius.
Pedestrian experiment	4	ST+TT	Two robots, each with ST model of self and ballistic TT model of other, demonstrates false beliefs.
Ethical robot experiments	5 & 6	ST+TT	One robot with ST model of self and ballistic TT model of one or two other robots. Demonstrates predictive modeling of self and reasoning about the intentions of others. Ballistic TT model extended so that the ethical robot can test and modify its belief about the proxy-human.
Imitation of goals	7	ST+ST	Imitating robot uses ST to model both itself and the demonstrator robot, in order to infer the demonstrator's goals.
Story-telling robots	8 & 9	ST+ST	Storytelling robot narrates *what-if* episode from its ST model; listener robot uses its ST model to introspectively 'imagine' that story. Potential to *explain* the past and possible future actions of self.

(2014) assert "Humans and robots, while vastly different, share a common problem of being embodied agents with sensory motor loops based on affecting and observing the physical world around them. By modeling a humans connection between mental states and the world as similar to its own, and reusing those mechanisms to help evaluate mental state consequences" a robot can at the basic level of actions and their consequences—model a human. The same is clearly also true for a robot modeling another robot of a different kind, providing that both observably sense and act in the physical world.

In the context of human-robot interaction we must consider the important problem of how a human builds a theory of mind for a robot; this could be especially important if that robot has the function of companion or elder-care (assisted living) robot. In the thought experiment of section 3.4 we outline how a robot's self-model can allow the robot to explain itself and hence assist a human to acquire an understanding of how and why the robot behaves in different circumstances.

The main contributions of this paper have been to (1) advance the hypothesis that simulation-based internal models represent a computational model of the simulation theory of mind (ST) and (2) to show that such a computational model provides us with a powerful and realizable basis for artificial theory of mind. We have shown that experiments with simulation-based internal models demonstrate the ability to predictively model the actions of both self and other agents. As summarized in **Table 1** the experiments of section 3 have demonstrated both ST+TT (hybrid) and ST+ST modes for self + other, as defined in section 2.3.

In summary, we contend that the experimental work outlined in this paper does demonstrate a number of components of theory of mind and can reasonably be described as "experiments in artificial theory of mind." The main hypothesis of this paper, that simulation-based internal modeling can form the basis for artificial theory of mind has, we argue, been demonstrated. Whilst far from a complete solution, we propose simulation-based internal modeling as a powerful and interesting starting point in the development of artificial theory of mind.

AUTHOR CONTRIBUTIONS

The author confirms that he is the sole creator of the text of this paper, and has approved it for publication.

FUNDING

The work of this paper is part funded by EPSRC grant ref EP/L024861/1, the German Academic Exchange Service (DAAD) and the Deutsche Forschungsgemeinschaft (DFG) grant HA 5458/9-1.

ACKNOWLEDGMENTS

The author is indebted to the outstanding research colleagues who conducted the experimental work outlined in this chapter, in particular Dr. Christian Blum, Dr. Wenguo Liu, Dr. Dieter Vanderelst, and Mathias Schmerling. The author is grateful to Susan Blackmore, Verena Hafner, Owen Holland, Susan Stepney, Richard Walsh, and the late Richard Gregory for many fruitful discussions on the topics in this paper. The author is also grateful to the reviewers of this paper and acknowledges the many improvements to the paper that followed.

REFERENCES

Astington, J., and Dack, L. (2008). "Theory of mind," in *Encyclopedia of Infant and Early Childhood Development*, eds M. M. Haith and J. B. Benson (Cambridge, MA: Academic Press), 343–356.

Birch, S., Li, V., Haddock, T., Ghrear, S., Brosseau-Liard, P., Baimel, A., et al. (2017). "Chapter 6: Perspectives on perspective taking: how children think about the minds of others," in *Advances in Child Development and Behavior*, Vol. 52, ed J. B. Benson (Cambridge, MA: Academic Press), 185–226.

Blum, C., Winfield, A. F. T., and Hafner, V. V. (2018). Simulation-based internal models for safer robots. *Front. Robot. AI* 4:74. doi: 10.3389/frobt.2017.00074

Bongard, J., Zykov, V., and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science* 314, 1118–1121. doi: 10.1126/science.1133687

Breazeal, C., and Scassellati, B. (2002). "Challenges in building robots that imitate people," in *Imitation in Animals and Artifacts*, eds K. Dautenhahn and C. L. Nehaniv (Cambridge, MA: MIT Press), 363–390.

Breed, M. D., and Moore, J. (2012). "Chapter 6: Cognition," in *Animal Behavior*, eds M. D. Breed and J. Moore (San Diego, CA: Academic Press), 151–182.

Brooks, R. A., Breazeal, C., Marjanović, M., Scassellati, B., and Williamson, M. M. (1999). "The cog project: building a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, ed C. L. Nehaniv (Berlin; Heidelberg: Springer), 52–87.

Carpenter, M., Call, J., and Tomasello, M. (2005). Twelve- and 18-month-olds copy actions in terms of goals. *Dev. Sci.* 8, 13–20. doi: 10.1111/j.1467-7687.2004.00385.x

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behav. Brain Sci.* 32, 121–138. doi: 10.1017/S0140525X09000545

Carruthers, P., and Smith, P. (eds.). (1996). *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Conway, M. A. (2005). Memory and the self. *J. Mem. Lang.* 53, 594–628. doi: 10.1016/j.jml.2005.08.005

Deák, G. O., Fasel, I., and Movellan, J. (2001). "The emergence of shared attention: using robots to test developmental theories," in *Proceedings 1 st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies*, eds C. Balkenius, J. Zlatev, H. Kozima, K. Dautenhahn, and C. Breazeal (Lund), 95–104.

Devin, S. and Alami, R. (2016). "An implemented theory of mind to improve human-robot shared plans execution," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16 (Piscataway, NJ: IEEE Press), 319–326.

Gariépy, J.-F., Watson, K. K., Du, E., Xie, D. L., Erb, J., Amasino, D., et al. (2014). Social learning in humans and other animals. *Front. Neurosci.* 8:58. doi: 10.3389/fnins.2014.00058

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York, NY: Oxford University Press.

Görür, O. C., Rosman, B., Hoffman, G., and Albayrak, S. (2017). "Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention," in *Workshop on The Role of Intentions in Human-Robot Interaction at 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)* (Vienna).

Gray, J., and Breazeal, C. (2014). Manipulating mental states through physical action. *Int. J. Soc. Robot.* 6, 315–327. doi: 10.1007/s12369-014-0234-2

Helbing, D., and Molnar, P. (1995). Social force model for pedestrian dynamics. *Phys. Rev. E* 51:4282.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends Cogn. Sci.* 6, 242–247. doi: 10.1016/S1364-6613(02)01913-7

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Res.* 1428, 71–79. doi: 10.1016/j.brainres.2011.06.026

Holland, J. (1992). Complex adaptive systems. *Daedalus* 121, 17–30. Available online at: http://www.jstor.org/stable/20025416

Holland, O. (eds.). (2003). *Machine Consciousness*. Exeter: Imprint Academic.

Holland, O., and Goodman, R. (2003). "Robots with internal models," in *Machine Consciousness,* ed O. Holland (Exeter: Imprint Academic), 77–109.

Kaplan, F., and Hafner, V. V. (2006). The challenges of joint attention. *Interact. Stud.* 7, 135–169. doi: 10.1075/is.7.2.04kap

Kennedy, W. G., Bugajska, M. D., Harrison, A. M., and Trafton, J. G. (2009). "Like-me" simulation as an effective and cognitively plausible basis for social robotics. *Int. J. Soc. Robot.* 1, 181–194. doi: 10.1007/s12369-009-0014-6

Kim, K.-J., and Lipson, H. (2009). "Towards a simple robotic theory of mind," in *Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '09 (New York, NY: ACM), 131–138.

Koenig, N., and Howard, A. (2004). "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.(IROS 2004)* (IEEE), 2149–2154.

Liu, W., and Winfield, A. F. T. (2011). Open-hardware e-puck Linux extension board for experimental swarm robotics research. *Microprocess. Microsyst.* 35, 60–67. doi: 10.1016/j.micpro.2010.08.002

Marques, H., and Holland, O. (2009). Architectures for functional imagination. *Neurocomputing* 72, 743–759. doi: 10.1016/j.neucom.2008.06.016

Michel, O. (2004). Webots: professional mobile robot simulation. *Int. J. Adv. Robot. Syst.* 1, 39–42. doi: 10.5772/5618

Michlmayr, M. (2002). *Simulation Theory Versus Theory Theory: Theories Concerning the Ability to Read Minds*. Master's thesis, Leopold-Franzens-Universität Innsbruck.

Mondada, F., Bonani, M., Raemy, X., Pugh, J., Cianci, C., Klaptocz, A., et al. (2009). "The e-puck, a robot designed for education in engineering," in *Proceedings of the 9th Conference on Autonomous Robot Systems and Competitions* (Castelo Branco: IPCB, Instituto Politécnico de Castelo Branco), 59–65.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018). Machine theory of mind. *arXiv:1802.07740*.

Roberts, W. (2001). "Animal cognition," in *International Encyclopedia of the Social and Behavioral Sciences*, eds N. J. Smelser and P. B. Baltes (Pergamon), 500–505.

Rohmer, E., Singh, S. P. N., and Freese, M. (2013). "V-rep: a versatile and scalable robot simulation framework," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Tokyo), 1321–1326.

Ruffman, T., Slade, L., and Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Dev.* 73, 734–751. doi: 10.1111/1467-8624.00435

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Auton. Robots* 12, 13–24. doi: 10.1023/A:1013298507114

Sebastian, M. A. (2016). Consciousness and theory of mind: a common theory? *Theoria Revista de Teoria, Historia y Fundamentos de la Ciencia* 31, 73–89.

Stafford, R. Q., MacDonald, B. A., Jayawardena, C., Wegner, D. M., and Broadbent, E. (2014). Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. *Int. J. Soc. Robot.* 6, 17–32. doi: 10.1007/s12369-013-0186-y

Stepney, S., Polack, F. A. C., Alden, K., Andrews, P. S., Bown, J. L., Droop, A., et al. (2018). *Engineering Simulations as Scientific Instruments.* (Cham: Springer).

Terada, K., and Ito, A. (2010). "Can a robot deceive humans?" in *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, HRI '10 (Piscataway, NJ: IEEE Press), 191–192.

Vanderelst, D., and Winfield, A. (2016). The dark side of ethical robots. *arXiv:1606.02583*.

Vanderelst, D. and Winfield, A. (2017). Rational imitation for robots: the cost difference model. *Adapt. Behav.* 25, 60–71. doi: 10.1177/1059712317702950

Vanderelst, D. and Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cogn. Syst. Res.* 48, 56–66. doi: 10.1016/j.cogsys.2017.04.002

Vaughan, R. T., and Gerkey, B. P. (2007). "Really reused robot code from the player/stage project," in *Software Engineering for Experimental Robotics*, ed D. Brugali (Berlin; Heidelberg: Springer-Verlag), 267–289.

Vaughan R., and Zuluaga M. (2006). "Use your illusion: sensorimotor self-simulation allows complex agents to plan with incomplete self-knowledge," in *From Animals to Animats 9. SAB 2006. Lecture Notes in Computer Science,* Vol. 4095, eds S. Nolfi, G. Baldassare, R. Calabretta, J. Hallam, D. Marocco, O. Miglino, J.-A. Meyer, and D. Parisi (Berlin; Heidelberg: Springer), 869.

Wagner, A. R., and Arkin, R. C. (2011). Acting deceptively: providing robots with the capacity for deception. *Int. J. Soc. Robot.* 3, 5–26. doi: 10.1007/s12369-010-0073-8

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

Winfield, A. (2014). Robots with internal models: a route to self-aware and hence safer robots.

Winfield, A. F. (2010). "You really need to know what your bot(s) are thinking about you," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed Y. Wilks (Amsterdam: John Benjamins), 201–208.

Winfield, A. F. (2018). "When robots tell each other stories: the emergence of artificial fiction," in *Narrating Complexity*, eds S. Stepney and R. Walsh (Cham: Springer), 1–11.

Winfield, A. F., and Hafner, V. V. (2018). "Anticipation in robotics," in *Handbook of Anticipation: Theoretical and Applied Aspects of the Use of Future in Decision Making*, ed R. Poli (Cham: Springer), 1–30.

Winfield A. F. T., Blum C., and Liu, W. (2014). "Towards an ethical robot: internal models, consequences and ethical action selection," in *Advances in Autonomous Robotics Systems, Vol. 8717, TAROS 2014. Lecture Notes in Computer Science,* eds M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish (Cham: Springer), 284.

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., et al. (2018). The grand challenges of science robotics. *Sci. Robot.* 3:eaar7650. doi: 10.1126/scirobotics.aar7650

Zagal, J., Delpiano, J., and Ruiz-del Solar, J. (2009). Self-modeling in humanoid soccer robots. *Robot. Auton. Syst.* 57, 819–827. doi: 10.1016/j.robot.2009.03.010

Zagal, J. C., and Lipson, H. (2009). "Resilient behavior through controller self-diagnosis, adaptation and recovery," in *Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '09 (New York, NY: ACM), 139–146.

Conflict of Interest Statement: 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.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 13 August 2018
doi: 10.3389/frobt.2018.00088

Toward Self-Aware Robots

Raja Chatila[1]*, *Erwan Renaudo*[1,2], *Mihai Andries*[1,3], *Ricardo-Omar Chavez-Garcia*[1,4], *Pierre Luce-Vayrac*[1], *Raphael Gottstein*[1], *Rachid Alami*[2], *Aurélie Clodic*[5], *Sandra Devin*[2], *Benoît Girard*[1] and *Mehdi Khamassi*[1]

[1] *Institute of Intelligent Systems and Robotics, Sorbonne Université, CNRS, Paris, France,* [2] *Intelligent and Interactive Systems, Department of Computer Science, University of Innsbruck, Innsbruck, Austria,* [3] *Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal,* [4] *Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Università della Svizzera Italiana - Scuola universitaria professionale della Svizzera italiana (USI-SUPSI), Lugano, Switzerland,* [5] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*

Despite major progress in Robotics and AI, robots are still basically "zombies" repeatedly achieving actions and tasks without understanding what they are doing. Deep-Learning AI programs classify tremendous amounts of data without grasping the meaning of their inputs or outputs. We still lack a genuine theory of the underlying principles and methods that would enable robots to understand their environment, to be cognizant of what they do, to take appropriate and timely initiatives, to learn from their own experience and to show that they know that they have learned and how. The rationale of this paper is that the understanding of its environment by an agent (the agent itself and its effects on the environment included) requires its self-awareness, which actually is itself emerging as a result of this understanding and the distinction that the agent is capable to make between its own mind-body and its environment. The paper develops along five issues: agent perception and interaction with the environment; learning actions; agent interaction with other agents—specifically humans; decision-making; and the cognitive architecture integrating these capacities.

Keywords: self-awareness, affordance, human-robot interaction, cognitive architecture, learning, decision-making, planning, Markovian processes

OPEN ACCESS

Edited by:
Antonio Chella,
Università degli Studi di Palermo, Italy

Reviewed by:
Danilo Emilio De Rossi,
Università degli Studi di Pisa, Italy
Owen E. Holland,
University of Essex, United Kingdom

****Correspondence:***
Raja Chatila
Raja.Chatila@isir.upmc.fr

Specialty section:
This article was submitted to
Humanoid Robotics,
a section of the journal
Frontiers in Robotics and AI

Received: *01 February 2018*
Accepted: *03 July 2018*
Published: *13 August 2018*

Citation:
Chatila R, Renaudo E, Andries M, Chavez-Garcia R-O, Luce-Vayrac P, Gottstein R, Alami R, Clodic A, Devin S, Girard B and Khamassi M (2018) Toward Self-Aware Robots. Front. Robot. AI 5:88. doi: 10.3389/frobt.2018.00088

1. INTRODUCTION

We are interested here in robotic agents, i.e., physical machines with perceptual, computational and action capabilities. We believe we still lack a genuine theory of the underlying principles and methods that would explain how we can design robots that can understand their environment and not just build representations lacking meaning, to be cognizant about what they do and about the purpose of their actions, to take timely initiatives beyond goals set by human programmers or users, and to learn from their own experience, knowing what they have learned and how they did so.

1.1. Context And Related Work

These questions are not new. Researchers in cognitive science, neurosciences, artificial intelligence and robotics have addressed the issues of the organization and operation of a system (natural or artificial) capable of performing perception, action, deliberation, learning and interaction, up to different levels of development (Morin, 2006).

The term "cognitive architectures" is commonly used in the Cognitive Sciences, Neuroscience and Artificial Intelligence (AI) communities to refer to propositions of systems organization models designed to model the human mind. Among the most renown long-term projects that propose cognitive architectures with the purpose of generality, two are particularly relevant to mention in the present context: (1) The SOAR architecture, standing for *State, Operator And Result*, proposed by Lehman et al. (2006); (2) the ACT-R architecture, standing for *Adaptive Control of Thought-Rational* proposed by Anderson et al. (2004). SOAR aims at modeling human cognition and is based on Alan Newell's seminal work on theories of cognition (Newell, 1990). Operational knowledge in SOAR is represented by production rules. To achieve a goal, the rules conditions are matched to a "working memory," of which the contents is encoded as sets of attribute-values. Learning in SOAR is mainly based on a mechanism called "chunking" (other mechanisms such as reinforcement learning are being added). This process is similar to identifying macro-operators, i.e., new rules that abstract the succession of rules selected to achieve a goal.

The general concept in ACT-R is a classical rule-based system. Knowledge about facts and events and their relationships is organized in a declarative memory along with a set of production rules and procedures. The memory component contains data structures called "chunks" whose meaning is nevertheless quite different from the chunks used in SOAR. The rules associated to selecting particular chunks depend first on the existence of matching elements in memory, and second also depend on the estimated probability of success and cost of their execution. Applying these rules can result in two different operations: either trigger robot action in the world, or change the corresponding elements in declarative memory. Each chunk in memory is also associated to a "base level" which increases proportionally to the number of times they have been selected in the past. This results in using chunks that have already been selected, i.e., that were used in more successful activations of the rules. The costs and success rates of the rules is modified according to the outcome of their execution. This leads to an improvement of the global behavior through time. Furthermore, there is a "compilation" process that produces new rules from analyzing the chunks involved in goal achievement.

These two major cognitive architectures present numerous common points. First, they both employ symbolic representations at high levels of abstractions. Second, they both use production rules to represent operational knowledge. Learning mechanisms in both architectures is mainly based on a memory of the success associated to prior action execution. Neither of these architectures really tackle the issue of operating in real time, nor the issue of how to build novel internal representations from sensory data. In practice, the authors of both architectures say that these are important issues, but no clear approach is put forward to overcome these issues. Another important issue is how to link symbolic and sub-symbolic representations, which goes beyond these proposals. Nevertheless, for applications to robots operating in real-time in the world, perceiving and manipulating unprepared sensory data, this question is central.

Most previous research aiming at developing robot cognitive architectures did not address the issue of self-awareness, an expression of consciousness which is a notion that requires to be clarified, whose foundations are not proven, and which is even considered as an illusion by some neuroscientists (Hood, 2012), while others propose to ground it in the solid theoretical framework of Integrated Information Theory (Koch et al., 2016).

We want to investigate if and how a machine can develop self-awareness. By doing so, we aim at understanding the concept itself and to propose computational models that can account for it (Chella and Manzotti, 2007; Lewis et al., 2011). The paper describes how the notion of self-awareness could be related to the development and *integration* of perceptual abilities for self-localization and environment interpretation, decision-making and deliberation, learning and self-assessment, and interaction with other agents. Such an integration appears to be key to enable the robot to develop some sense of *agency*, or the awareness of being in control of its own actions and responsible for their outcome (Haggard and Tsakiris, 2009). Moreover, such an integration of the results and characteristics of various subconscious deliberative processes (such as perception, action and learning) in a common *global workspace* (Dehaene and Naccache, 2001) appears fundamental in humans to enable meta-cognitive processes such as the ability to report to oneself and to other agents about her internal state, her decisions and the way these decisions were made (Shadlen and Kiani, 2011), but also importantly to develop predictive models of agency (Seth et al., 2012).

The processes implementing these capacities must operate simultaneously for online performance in robots interacting in real-time with their environment as well as with other agents. Furthermore, central to this project is the design of an architecture that constitutes a robotic model of an efficiency-based performance testbed for the integration of these processes, and which could in a second stage be used to qualitatively (and even maybe quantitatively Oizumi et al., 2014) assess the emergence of minimal degrees of awareness as a result of their interaction for the resolution of a set of tasks. Our goal is to explore this assertion and to demonstrate it with experimental proofs of concepts.

The rationale of this paper is that the understanding of its environment (including other agents) by an agent requires its self-awareness, which actually is itself emerging as a result of this understanding and the distinction that the agent is capable to make between its own mind-body and its environment. This constitutes a dynamical system in which some authors have proposed that the awareness of self through stability and distinctiveness can be built (Marks-Tarlow, 1999; Shoda et al., 2002). We claim moreover that on the road toward a better understanding of the integration mechanisms underlying awareness, the successes and failures of robotics investigations can be useful in identifying what is not awareness, for instance when exemplifying some robotic *zombies* which can solve without awareness tasks that are thought to involve awareness (Oizumi et al., 2014).

1.2. What Is Self-Awareness?

We will not attempt a strong definition of self-awareness, but we try in this paper to ground the concept. Our hypothesis is that self-awareness must first rely on perception of self as different from the environment and from other agents. This necessitates that the robot interacts with the environment and build sensory-motor representations that express the affordance of environment elements to it, and that it interacts with other agents to distinguish itself from them. Affordance building is presented in section 2 and distinction from and reasoning on other agents is discussed in section 4. Building on environment representations that integrate perception and action, two main capacities are introduced that we believe are necessary enabler of self-awareness:

- Self-evaluation. This is the capacity of "knowing that I know" and deliberately using this knowledge in action selection. I n other words, the robot builds a knowledge on what abilities it has learned and when it can use them. It is able to transform learnt behaviors into explicit skills and to characterize the situations in which these skills are applicable, reverting to a planned goal-directed behavior when they are not. This is presented in section 3.
- Meta-reasoning. The other main capacity is deliberation on one's own reasoning. In section 5, we propose a system initially driven by basic motivations, able to reason on the means for satisfying them to determine its own goals. Eventually, new motivations should be learned but this is not developed in the paper.

In section 6 the cognitive architecture for integrating all robot capacities is presented, but a validation of this global architecture still remains to be done. Finally we conclude in section 7.

2. PERCEPTION AND LEARNING AFFORDANCES

Traditionally, in robotics perception (excluding visual servoing and similar closed-loop control) is considered only as an isolated observation process. We believe that this approach undermines the capacity of current agents (i.e., robots) for scene understanding. Simultaneously perceiving and acting requires to interpret the scene with respect to the agent's own perceptual capacity and its potential activities. What an agent can do (or *afford*) with an object partly circumscribes the meaning that this object can have for her: a mug on a table is something that can be filled with liquid and then brought to the mouth in order to drink for a human; the same object is a place on which to land (and possibly eat) for a fly, behind which to hide for a mouse, or something that can be pushed to the ground producing a fancy noise for a child. This interpretation fits with Gibson's notion of affordance (Gibson, 1977; Sahin et al., 2007).

Reasoning jointly on perception and action requires self-localization with respect to the environment. Hence developing sensorimotor representations and not just exteroceptive representations puts the robot in the center of the perceptual process, and provides a link between self-awareness and situation-awareness. Robot localization with respect to its environment provides a differentiation between the robot's body and the external world, and includes a necessary distinction between its parts and surrounding objects. In addition, robot's *actual* components link robot's body-environment's state before and after actions are applied.

In this section we propose sensory-motor representations and scene interpretation processes that integrate four inputs: perceptual (perceiving the external scene), proprioceptive (input from the agent's own configuration), contextual (previous knowledge) and the agent's action capabilities.

We propose a methodology to build models of objects based on perceptual clues and effects of robot's actions on them. Our methodology employs a Bayesian Network for representing the robot's actions, the objects in the environment, as well as changes in the observable environment triggered by the robot's actions. We then perform structure learning on continuous and discrete variables representing these informations in order to identify the most probable Bayesian network that best fits the observed data. Analyzing the structure of the obtained Bayesian network permits the robot to discover correlations between itself and the environment using statistical data.

The proposed affordance learning architecture is depicted in **Figure 1**. Measurements from the *Environment Interaction* are the main inputs of our approach, it includes visual perception from camera and proprioception values from joints. A set of clusters are extracted from clouds of points through *Visual perception*. Clusters are then tracked to generate hypotheses about the objects the robot interacts with. Proprioceptive feedback is retrieved under the form of measurements of joint and force. Then the input from perception and action tasks is analyzed by *Effect detectors* to extract salient changes from the interaction process. At the intersection between the two input processes is *Sensory-motor learning* which represents the fusion between the perception and action components. *Affordances learning* process relates *objects*, *actions* and induced changes considered as *effects* to build the final sensory-motor representation. A *Motivational system* orchestrates the process of selecting objects and actions that will be applied on them. The final representation is saved in a *long-term storage* which also provides feedback to the motivational system.

While interacting with the environment, the robot infers dependencies between the affordance elements (*objects, actions*, and *effects*) thus combining perceptual and proprioceptual data. The robot's motivational system relies on the learned sensory-motor representations and the Beyesian framework to make predictions about a set of affordance elements. This inferred information can be used for learning decisions, for future planning tasks, or to add sensor and motor capabilities to the innate repertoire.

2.1. Exteroceptive Perception

To our knowledge, most existing segmentation algorithms mainly focus on raw to low level information from the 2D image or 3D point cloud. However, some recent methods for semantic segmentation have been proposed which can disambiguate object borders by taking advantage of high-level object knowledge (Silberman et al., 2012; van Hoof et al., 2014). However, the computational cost of inference on these methods rises considerably with the increasing number of objects. Moreover, the relations between nodes come from a

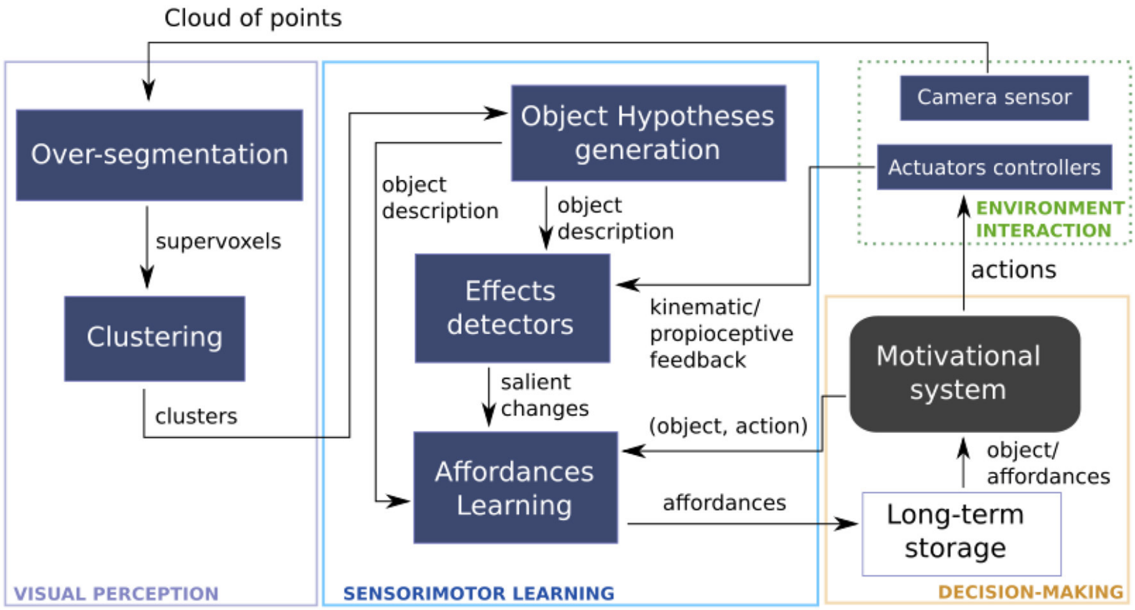


FIGURE 1 | Architecture of the proposed sensorimotor approach for scene affordance learning.

priori information from the objects class, which limits their use in self-discovered scenarios.

2.1.1. Over-Segmentation

Over-segmenting a color cloud of points into small regions based on local low-level features of geometry and color enables to form supervoxels. We implemented a 3D version of the Voxel Cloud Connectivity Segmentation (VCCS) (Papon et al., 2013), which generates evenly distributed supervoxels. VCCS employs a flow-constrained local iterative clustering process which uses geometric features and color, and a seeding methodology based on 3D space. The seeding of supervoxel clusters is done by partitioning 3D space to ensure that supervoxels are evenly distributed according to the geometry of the scene. Strict spatial connectivity of occupied voxels can be enforced by the iterative clustering algorithm. This algorithm guarantees that supervoxels cannot flow across boundaries which are disjoint in 3D space even if they are connected in the projected plane.

Supervoxels are represented by a 39-dimension feature vector composed of 33 elements from an extension of the Fast Point Feature Histogram (FPFH) (Papon et al., 2013), color information (Lab color space) and spatial coordinates (x, y, z). This permits to exploit a pose-invariant multi-dimensional representation based on the combination of neighboring points. **Figure 2** (middle) depicts an over-segmented cloud where each supervoxel (representing a segment) cannot cross over object boundaries that are not spatially adjacent in 3D space.

Supervoxels in **Figure 2** (middle) only represent individual patches. A clustering process is needed to group the supervoxels that possibly correspond to the same object. The non-parametric technique described in Comaniciu et al. (2002) was implemented to find the shape of object hypotheses based on the set of clustered supervoxels.

Figure 2 (right) shows the result of the clustering method as a set of labels $L_{hyp}(t)$ for a cluster of supervoxels that may represent objects in the current scenario.

The set of generated segments (section 2.1.1) are built only using the sensory data. This means that segmentation issues can appear in the form of incomplete, divided and false segments of real objects in the scenario. We overcome this issue by performing a tracking-by-detection approach which reduces the number of false positive segmentations (Chavez-Garcia et al., 2016b). In this approach, each object is represented by its centroid, which additionally offers a point of interaction in further interaction tasks.

2.2. Sensory-Motor Learning

Manipulating objects enables the robot to not only perceive information, but also and most importantly to learn sensory-motor correlations between the robot's basic actions A, the sensory inputs contained in the objects' descriptions O, and the salient changes represented by the effects E. The objective here is to learn from regularities in the occurrences of elements in O and E when an action $a_i \in A$ is triggered. While the robot is starting the learning from built-in actions, this process permits to progressively develop a representation of the environment captured by perception through object movement detection and proprioceptive feedback.

2.2.1. Objects

We make the assumption that the robot has prior perceptual capabilities that enable it to discretize the environment. These capabilities are related to the segmentation approach. The robot

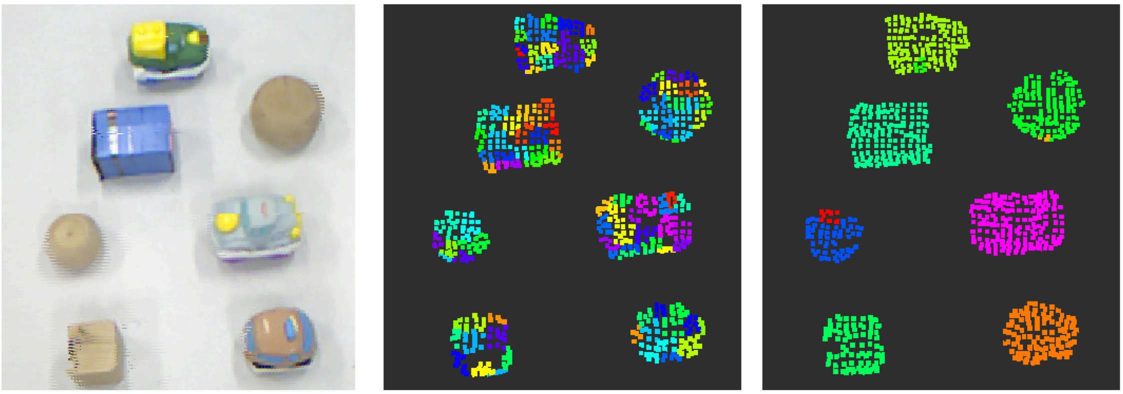


FIGURE 2 | Results from the perception process. Appearance and spatial information from the RGB-D point cloud of the real scene **(Left)**; supervoxels from over-segmentation of the point cloud **(Middle)**; and results from intrinsic clustering **(Right)**.

has prior geometrical notions of position, continuity of segments and normal extraction for surfaces, can recognize different color values, and using these perceptual capabilities can extract higher level features (e.g., as combinations) for describing confirmed objects. The cloud of points representing a object can provide relevant features, such as color, size and shape. Our architecture permits to incrementally learn the set of perceptual features which are relevant in the robot's surrounding environment.

2.2.2. Actions

We assume that the robot is built with a set of basic motor capabilities, or actions, described relative to the actor and its morphology. These basic actions $A = \{a_1, ..., a_n\}$ are defined with respect to their control variables in joint space:

$$a : \{Q, \dot{Q}, \ddot{Q}\}_\tau \tag{1}$$

where Q are the joint parameters of the robot used in action a, and τ the duration of this action. This implies that, by definition, two actors with completely different motor capabilities and morphologies cannot execute the same actions (but their effect might be identical).

The extraction of points of interest in the image representing a particular object is done by raising perceptual hypotheses about possible identifications of this objects. These points are used to reduce the set of possible actions that permit to approach the object through perceptual servoing. In that sense, the focus of our work is really on sensorimotor representation through object manipulation.

2.2.3. Effects

An effect is a correlation between an action and a change in the state of the environment, which includes the agent itself. Effect learning can be crucial to build internal world models used for learning and decision-making, consisting in actions' effect in terms of possible rewards and possible transitions to different states of the environment (see section 3 for examples of how the robot can use such world models).

When a robot interacts with an object it can perceive (via its exteroceptive capabilities) changes related to the position or appearance of the object, proprioceptive values from actuators and feedback from end-effectors. Effect detection (or lack thereof) represents the common ground for perception and action frames. Robot's capabilities to detect effects are divided into two groups: perceptual-based (e.g., changes in perceptual representations of objects); and proprioceptive-based (e.g., changes in robot's internal representations).

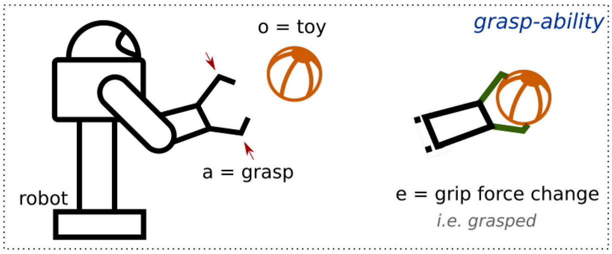


FIGURE 3 | Representation of the *grasp-ability* affordance relation.

2.2.4. Affordance Learning

We follow the definition of an affordance employed in Andries et al. (2018), where we consider O the set of objects, A the set of actions, and E the set of observable effects. When an actor g_m applies an action a_l on object o_k, generating the effect e_j, the corresponding affordance α is defined as:

$$\alpha = ((o_k, a_l), e_j), \text{ for } o_k \in O, a_l \in A \text{ and } e_j \in E, \tag{2}$$

This definition shows an affordance as an *acquired* relation between the elements in O, A, and E (Chavez-Garcia et al., 2016a).

An example of an affordance relation between the object *toy* and the *robot* is shown in **Figure 3**. It illustrates the application of the robot's capability *grasp*, implying that there is a potential of generating an effect *grasped* that can be detected by the robot's exteroceptive and proprioceptive capabilities (e.g., grip force change). Using the semantic value of this relation, we can label it as $grasp - ability$.

When the robot interacts with the environment, we record the values of each element in the affordances' sets. By considering each element as a random variable in a Bayesian network $\mathcal{B}$, the problem of discovering the relations between E, O and A can then be translated into finding dependencies between the variables in $\mathcal{B}$, i.e., $P(\mathcal{B}|\mathcal{D})$, which means learning the structure of the corresponding Bayesian network $\mathcal{B}$ from interaction data $\mathcal{D}$. In this way, affordances are described by the conditional dependencies between variables in $\mathcal{B}$.

The score of a structure is defined as the posterior probability given the data $\mathcal{D}$. We implemented an information-compression score that applies a penalization defined as $s(N) = \frac{log(N)}{2}$ to represent the number of bits needed to encode $\mathcal{B}$ (Chavez-Garcia et al., 2016b). This score penalizes structures with larger number of parameters.

We implemented a search-based structure learning algorithm based on the hill-climbing technique (Chavez-Garcia et al., 2016b). The algorithm receives as input the values of variables in E, O, and A recorded during robot's interactions. It attempts every possible single-edge addition, removal, or reversal, selecting as current top-candidate the network with the highest score, and iterating. For each tested structure the algorithm estimates the parameters of the corresponding local probability density functions. The process stops when the score can not be increased anymore by a single-edge change. Although this algorithm does not guarantee that it will settle on a global maximum, a simulated annealing technique was implemented to avoid getting stuck in local minima.

Such a robotic implementation of the Bayesian Network framework for perception allows the robot to display relationships between affordance elements. The directed nature of its structure approximates cause-effects relationships and includes uncertainty from the interaction process. Moreover, in addition to direct dependencies, the model can represent indirect causation. These elements are key to enable a first minimal level of self-awareness of the robot by being able to monitor the effects of its actions on the environment, differentiate itself, other agents, movable objects and fixed elements of the environment. The uncertainty about the learned effect can moreover enable the robot to display some degree of confidence about the things it learned and to explicitly require more interactive experience with the objects and actions for which it is less confident. Finally, the estimated transitions between states of the environment that can be learned within world models enable some degree of anticipation, permitting the robot to predict future states of the world depending on its actions and on the actions of the others (as we illustrate in the joint action framework presented in section 4). These capacities will be crucial for planning and model-based learning abilities developed in the next section.

3. LEARNING ACTIONS AND PLANS

One of the main points presented in this section is that the ability to coordinate different strategies for decision-making and reinforcement learning (here considered as the main adaptation process of decision-making) can constitute a first step toward (i) more robotic autonomy and adaptation, but also toward (ii) the capacity for the robot to analyse the efficiency of its decision-making processes and use this analysis to change not only its behavior but the way it generates its behavior. Moreover, performing efficient online dynamic coordination of multiple learning and decision-making systems requires the implementation of a *meta-controller* within the robot cognitive architecture, which observes what each system does, and predicts and monitors their effect on the robot's internal state and environment. This can thus participate further to the emergence of self-awareness as integration of deliberative and reporting processes.

Here we consider that the motivational system of the robot (see section 5) provides reward to the latter when it fulfills certain tasks (e.g., recharging its batteries in a particular location, or answering a human request). We further make the assumption that for the duration we consider, this motivation will remain stable. In order to accomplish the task and satisfy its motivation, the robot needs to act in its environment. Its action selection mechanisms are then in charge of producing the relevant behavior to reach the task's goal. These action selection mechanisms have been traditionally modeled by the robotic community by action planners (see Khamassi et al., 2016; Ingrand and Ghallab, 2017 for recent reviews). Planners produce a sequence of actions to bring the robot from its current state to the goal state. Initially based on first-order logic (Fikes and Nilsson, 1971), these planners have been extended with probabilistic methods to take into account uncertainty by modeling the problem as a Markov Decision Process (sometimes Partially Observable if the uncertainty is on states). This also allows to use reinforcement learning (RL) algorithms (Sutton and Barto, 1998) to find relevant policies.

In RL, two main categories of methods can be used: model-based methods learn and use the transition and reward models of the problem (respectively the structure of the state-action-state space and the reward signals in the state-action space); model-free methods locally learn the reward-predictive value associated with each state-action pair without explicitly taking into account the effects of the action predicted by a world model of the task. The former are comparable to planning, as they find the optimal policy (i.e., the best action plan) through a costly computation using a model of the task, and hence completely update the policy between two interactions with the environment. The latter are reactive methods allowing fast action selection but are slow to learn, requiring multiple interactions with the environment to locally update each state-action value. Each type of action selection process has its advantages and has been used in a variety of applications (Kober et al., 2013). However, research in robotics have only recently started to consider the possibility of combining these two different learning methods as parallel alternative strategies to solve the same task (Caluwaerts et al., 2012; Renaudo et al., 2014).

These multiple action selection systems architectures for robotics are inspired by biological evidence of a comparable systems-combination process in mammals. Neurobiological studies have highlighted the existence of a *goal-directed behavior* when mammals are moderately trained on an instrumental task (Yin and Knowlton, 2006; Dayan, 2009). This behavior is characterized by a decision-making process oriented toward an explicit goal representation. It is moreover hypothesized to rely on the progressive learning of an internal model of the task structure, the use of this model for prospective inference and planning being experimentally observable through transient increases in subjects' deliberation time (Viejo et al., 2015). This enables a high flexibility in response to sudden changes in the task (e.g., the source of reward is moved), because behaviors that the internal model do not estimate as leading to the goal anymore can be inhibited. On the other hand, extensive training in a familiar task makes the behavior *habitual*, which is illustrated by an increase in subjects' action rate and an insensitivity to task changes (Balleine and O'Doherty, 2010), in the same manner as one could persist with the sequence of finger presses corresponding to an old pin code after this code has been recently changed. Interestingly, while healthy mammals can switch back to goal-directed behavior after a short persistence time following a task change, lesions to different brain regions can either prolong or reduce this persistence period, thus suggesting that both types of behaviors might coexist and compete for control within a modular brain architecture (Yin and Knowlton, 2006).

While goal-directed and habitual behaviors have been modeled respectively as model-based and model-free RL algorithms (Daw et al., 2005), the question of the mechanisms underlying their coordination is still an active area of research in computational neuroscience (e.g., Viejo et al., 2015; Dollé et al., 2018). Nevertheless, here we do not investigate how to operationalize this coordination and to adaptively switch from model-based to model-free control with such a bio-inspired multiple action selection system architecture, because this has been the subject of our prior work (Renaudo et al., 2014, 2015b,c). Instead, we focus here on how such an architecture enables the robot to self-monitor these action selection systems, when they are advantageous and what advantage they bring (e.g., efficiency vs. rapidity), and thus how the robot can get the ability to self-report about the way it makes decisions while learning a particular task. This ability to self-monitor can be related to the notion of self-awareness and is stated as important to allow flexible and adaptive control of a being (Van Gulick, 2017).

3.1. Multiple Action Selection Systems Architecture

3.1.1. Overall Architecture

The architecture is presented in **Figure 4**. Each module (or *expert*) is a decision-making system that implements one way of producing actions: the goal-directed expert in a model-based RL manner and the habitual expert in a model-free RL manner. These experts learn either a model of the task or only the local state-action values based on the reward received from the motivational system and the experienced states and actions. States are received from robot sensor data processing and a set of discrete actions is made available to the action selection systems.

Whereas only one decision-making system (*expert* here) is sufficient for a robot to act autonomously, our architecture also integrates an additional component in charge of monitoring

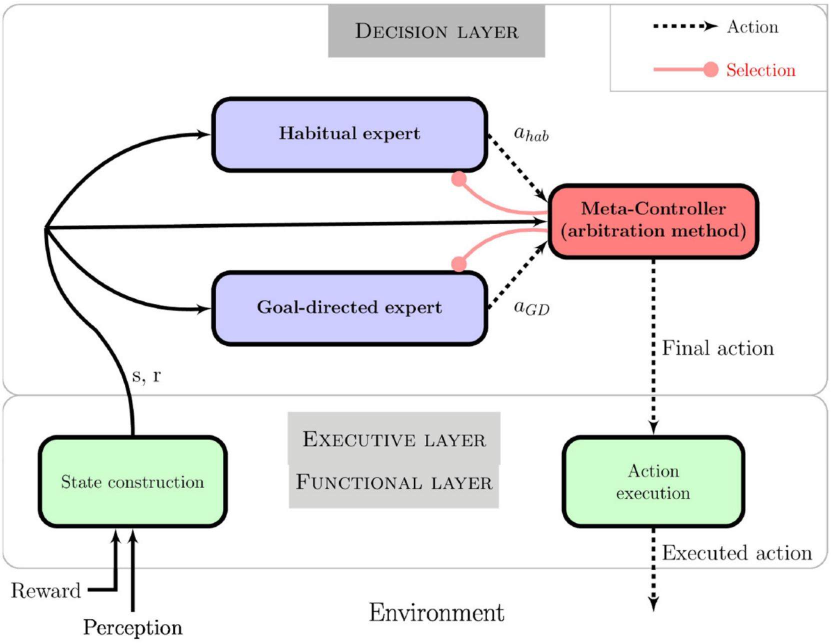


FIGURE 4 | Global action selection architecture composed of two decision systems implementing corresponding behaviors: the goal-directed expert is a model-based RL algorithm whereas the habitual expert is a model-free RL algorithm. The meta-controller is in charge of monitoring different expert information, giving control to one of the two. The reward information comes from the motivational system and represents the goal of the task.

the decision-making process. The *meta-controller* analyses each expert and selects which one is actually controlling the robot at a given time step. It implements the arbitration method studied hereafter. We argue that this component is necessary to allow the robot not only to act according to the task to be fulfilled, but also to criticize and report on its own decision process.

3.1.2. Possible Coordination Methods

In previous work (e.g., Renaudo et al., 2015b,c) we have studied and compared arbitration methods that can be separated in two categories: (i) fusion methods merging action selection probability distributions from each expert into a given state, and select an action from the final distribution and (ii) selection methods evaluating which expert is the most relevant in the current situation and let it decide about the final action. We have also defined a reference coordination method where each expert E among N experts ($N = 2$ here) has a constant and uniform probability $P(E)$ of being selected: $P(E) = 1/N = 0.5$ in this case. This random selection has been used as a proof-of-concept in earlier work and defines the bottom performance to evaluate the interest of each particular coordination method (Renaudo et al., 2014).

Comparison of these different tested coordination methods suggests that the arbitration method should take into account multiple signals rather than only one that will miss some of the required information (Renaudo et al., 2014, 2015c). It also suggests that arbitration and expert selection should rely on information available before the experts actually compute the action to perform in the current state: this allows to save computation time of the overall decision process.

Moreover, in previous similar works (Dollé et al., 2010; Caluwaerts et al., 2012; Dollé et al., 2018), the coordination is mostly performance-based: the meta-controller in these algorithms learns which expert to recruit in each state of the task in order to maximize reward, but does not consider each expert's specific properties. Here, the habitual expert is computationally less expensive than the goal-directed expert. Thus, in case of equal performance of the experts, self-monitoring these processes should allow the meta-controller to prefer the less costly expert. On the other hand, the goal-directed expert is more efficient to update the whole policy between two interactions with the environment. When the meta-controller observes that the habitual expert proposes irrelevant actions, it can decide to select the goal-directed expert despite its high computational and time costs.

Thus, to illustrate the interest of the self-monitoring capability provided by the meta-controller, we propose a new *Learning and Cost* arbitration method described hereafter.

3.1.3. A Coordination Method Based on *Learning and Cost* Signals

Building on these previous conclusions, we propose a new signal-based method that uses two measures of expert's status. Only the selected expert estimates the action values, which allows to save computation time and to be more reactive. One signal is directly related to this goal: the intrinsic computation cost incurred by each expert to evaluate action values. The other signal measures the experts' knowledge about the task, which can be evaluated by their learning progress.

We define $\overline{T}_{Hab}, \overline{T}_{GD}$ as the mean computation times for the two experts, evaluated with exponential moving averages (see Equation 3; $\lambda = 0.02$ which is equivalent of averaging over 50 decision steps). These means are updated only when their expert has been selected to make a decision, as no cost can be measured otherwise.

$$\bar{s}_t = (1 - \lambda) \cdot \bar{s}_{t-1} + \lambda \cdot s_t \tag{3}$$

We define $\overline{\delta Q}$ as the mean variation of Q-values reflecting the progress of learning in the habitual expert, and $\overline{\delta P}$ as the mean variation of the transition model probabilities reflecting the progress of learning in the goal-directed expert. In model-based RL, learning is about estimating the task's transition and reward functions. Thus a measure of learning progress should refer to the model's estimation rather than the computed Q-values. The mean variations are updated after each action with an exponential moving average ($\lambda = 0.2$ or 5 decision steps).

In order to combine cost and learning information, we define V_E, the *value of selecting expert E* as the weighted sums in Equation (4). We seek to preferentially select the expert that computes at the lowest cost, and that does not need to update much its knowledge because it already has enough information about how to solve the task:

$$\begin{aligned} V_{Hab} &= -(\alpha_{Hab} \cdot \overline{\delta Q} + \beta_{Hab} \cdot \overline{T_{Hab}}) \\ V_{GD} &= -(\alpha_{GD} \cdot \overline{\delta P} + \beta_{GD} \cdot \overline{T_{GD}}) \end{aligned} \tag{4}$$

The α_i and β_i parameters are the positive weights of each signal in the selection. As $\overline{\delta P}$ and $\overline{\delta Q}$ have different amplitude ranges, we set $\alpha_{GD} = 1$ and $\alpha_{Hab} = 12$, so the transition from goal-directed expert to habitual expert needs a strong convergence of Q-values in the model-free algorithm. $\beta_{Hab} = \beta_{GD} = 5$ in order not to bias the selection and to keep the natural difference in expert costs. Since the GD expert is computationally more costly than the Hab one, this method makes the meta-controller preferentially select the latter more often when the learning progress is equivalent between experts. These values are converted into selection probabilities $P(E)$ using a softmax function (5) from which the selected expert is drawn. As expert E pays the cost of estimating actions only if it is selected, its corresponding $\overline{T_E}$ is only updated in the latter case.

$$P(E) = \frac{\exp(V_E/\tau)}{\sum_{b \in \mathcal{A}} \exp(V_E/\tau)} \tag{5}$$

In this method, τ is set to 1.

3.1.4. Evaluation in a Navigation Task

We evaluated the approach of combining multiple action selection systems in simulation in previous work. Especially, preliminary analyses of the reference method in a simulated human-robot interaction task (see **Figure 5**, left) have been

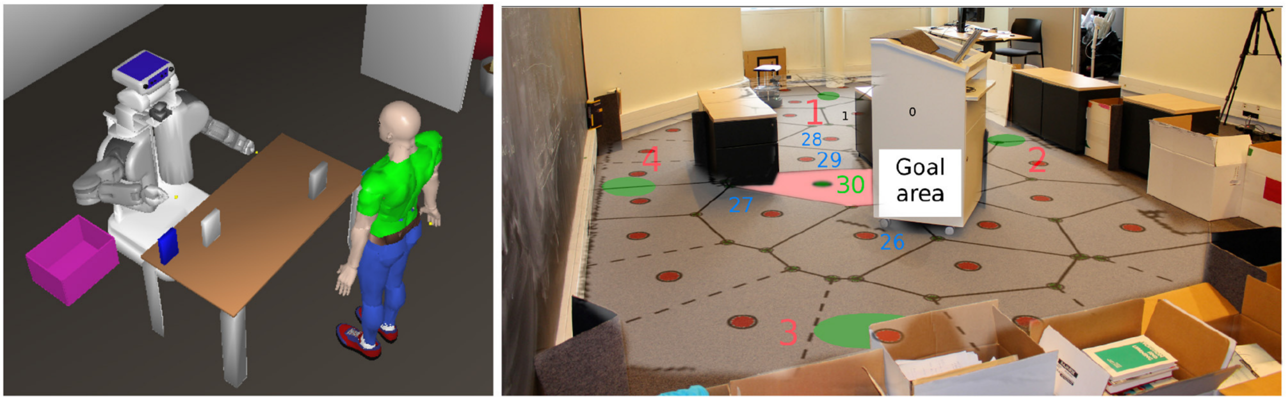


FIGURE 5 | (Left) Setup for the Human-Robot Interaction (HRI) task from Renaudo et al. (2015a): the human and the robot collaborate to put all boxes in a trashbin . **(Right)** Arena for the navigation task. A mapping of the states produced by the robot has been manually added. The red area indicates the goal location whereas the green areas indicate starting locations of the robot. Red numbers are starting location indexes; blue numbers are some states indexes referred to later.

reported in Renaudo et al. (2015a) and are further discussed in the next section on human-robot interaction. Here, we present novel results with the *Learning and Cost* method applied to a real robot in a navigation task.

In this task, a Kobuki Turtlebot robot has to navigate from starting locations (see **Figure 5**, right, green areas numbered 1–4) to the center of a 7.5 m × 3.5 m arena. Two obstacles split the arena in three corridors, the goal being located in the middle one (red area). The reward (1 unit) is given when the robot enters the goal area. It is then driven back to one of the starting locations (randomly selected). The robot localizes itself thanks to a standard particular filter based SLAM algorithm (Grisetti et al., 2007). The occupancy map built by exploring the environment is discretized into about 30 states following a regular paving. In each state, the robot can select between the 8 directions around it in the world frame. The robot controller takes care of driving the robot in the chosen direction and avoiding obstacles. We evaluate again three configurations: (i) goal-directed expert alone, (ii) habitual expert alone, (iii) both experts operate (*Combo*) and are coordinated by the meta-controller with the *Learning and Cost* method. Each configuration is evaluated 10 times, the habitual expert alone is given 2 h per repetition to learn from scratch, the goal-directed expert alone and the combination are given 1h but benefit from 1h of latent exploration (without reward in the environment) to allow the goal-directed expert to build its transition model.

3.2. Results

The first result of this experiment confirms the results from previous work. **Figure 6** shows the final weights (which are direct images of the *Q*-values) of the habitual expert in states near the goal. When the latter is controlling the robot alone, learning is long and the *Q*-values are weakly discriminating which action will give the highest reward. When control is shared with the goal-directed expert according to the *Learning and Cost* method, the habitual expert learns faster (mostly bootstrapped through observation of the behavior produced by the GD expert), which is represented by more contrasted final values in these states.

Figure 7 shows the monitored signals during the navigation task. Time 0 represents the initialization of a new goal location. Not surprisingly, the cost of using the goal-directed expert is one order of magnitude higher than the habitual expert cost. Interestingly, during the first minutes the habitual expert is more often selected than the GD expert until the new goal location is discovered and the GD expert starts making less error so that it gets more selected by the meta-controller. Then starts a long habit learning phase where the Hab expert slowly learns the new appropriate state-action values, which penalizes its selection (due to the high value of a_{Hab} in the criterion). As the two experts are in different states of knowledge on how to perform the task, the meta-controller mostly selects the goal-directed expert, certainly more costly but more reliable to produce the best behavior.

Here, given the real robotic setup and the natural slow learning speed of the habitual expert, the control goes mainly to the goal-directed expert. In different conditions or with longer time, the *Q*-values of the habitual expert can stabilize and this method favors its selection. Nevertheless, the important message here relative to the issue addressed in this paper is that these monitoring signals can be used by the robot to analyse its own decision-making processes and evaluate which decision-making strategies (GD or Hab) were the most efficient at different phases of the task. These capacities to monitor and report about its own performance can be integrated with representations of other agents' own abilities for efficient joint action.

4. HUMAN-ROBOT INTERACTION: AGENT AWARE TASK PLANNING

For more than a decade, the field of human-robot interaction has generated many valuable contributions of interest to the Robotics community at large. We will here give some insights

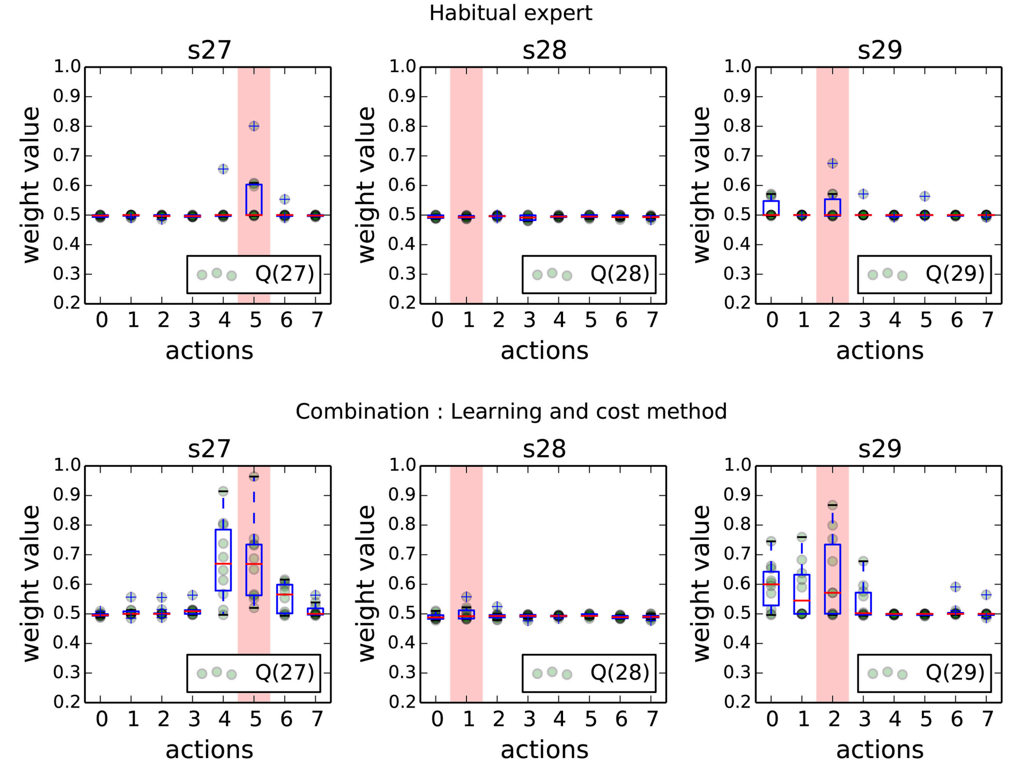


FIGURE 6 | Weights of each action (direct image of Q-values) for the habitual expert when alone **(Top)** or combined with the goal-directed expert **(Bottom)** at the end of the navigation task. Each light green dot is the final learned value of each action. The red bar indicates the best action to take from the human perspective. These measures are shown in the states next to the goal (s27, s28, s29).

concerning a particular type of interaction which is joint action, and the associated required levels of awareness. To do so, we will first explain which processes are involved in human-human joint action and then in human-robot joint action, in order to argue that minimal levels of self-awareness are required for the robot to efficiently integrate information about the effects of its actions and the effects of other agents' actions into feasible joint action plans.

4.1. Human-Human Joint Action

In order to establish successful joint action, interacting agents need to be able to efficiently share and coordinate their intentions, plans, goals and actions with other participants. Put it differently, it is not enough to share a common goal between interacting agents to establish efficient joint action if each agent then individually chooses his/her own sub-goals, and simply devise his/her own individual action plan and executes it. There is a need to share a coherent joint action plan but also to coordinate actions and sub-plans between agents. This coordination is particularly crucial during the execution phase in order to ensure the successful completion of the joint action (Clark, 1996; Grosz and Kraus, 1996; Bratman, 2014; Clodic et al., 2017). One possible way to do that is from the point of view of each agent to monitor both his/her own actions and intentions as well as those of his/her partner's. Such a monitoring process can facilitate the representation and understanding of the combined impact of agents' actions on their shared goal, and the adjustment of what they do accordingly.

An important ingredient of this agent coordination process which goes in complementarity with the co-representation of tasks and actions is *joint attention*. It is an ability that has been found in apes to provide a key mechanism for establishing common ground in joint action by sharing perceptual representations of the surrounding environment and task such as the available objects and the occurring events (Tomasello and Carpenter, 2007). As an example, Brennan et al. (2008) had participants engage in a joint visual search task and showed that they were able to most of the time focus on a common space between them by directing their attention toward portions of the environment where the other was looking. Moreover, they found that their performance during such a joint search task was improved compared to the one obtained in an individual version of the task. Besides, Vesper (2014) have shown that co-agents not only engage in joint attention but also repeatedly perform transient modulations of their own movements that "reliably [have] the effect of simplifying coordination." These are known as *coordination smoothers* and are part of a more general process called *signaling* which constitutes another phenomenon that contributes to better on-the-fly coordination. A particularly striking example is when someone exaggerates his/her own movements or reduces

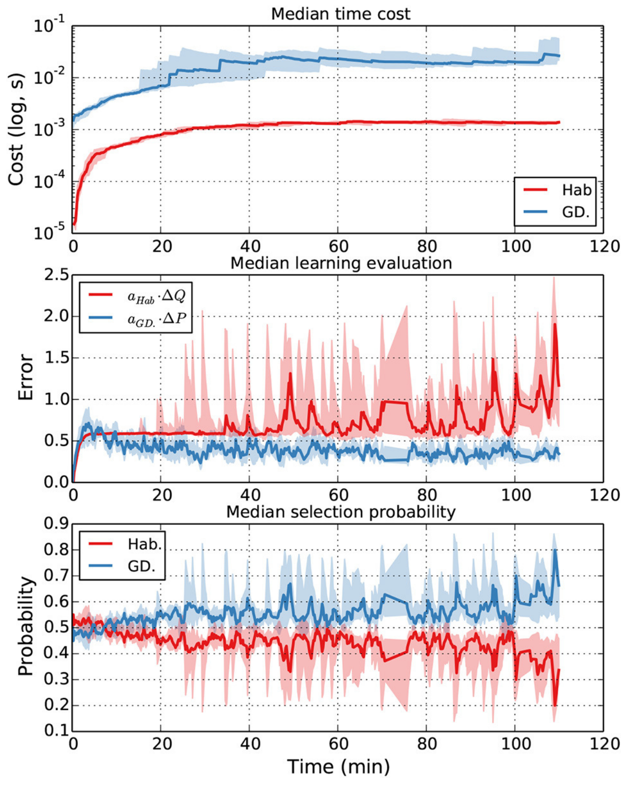


FIGURE 7 | Evolution of monitored signals when both experts are controlling the robot during the navigation task. **(Top row)** shows the sliding mean cost spent by both experts for decision-making. **(Middle row)** shows the measures of learning scaled by their coefficient. **(Bottom row)** shows the evolution of the probability of selection of each expert. In these experiments, the strong parameter of the habitual expert learning measure combined with its slow convergence favors the goal-directed selection in order to reach the goal more easily (however at a high computational cost).

his/her movement variability in order to make them more easily understandable and interpretable by the other participant (Pezzulo et al., 2013).

It is important in contrast to take into account any form of joint action that may not require awareness. For instance, perception-action couplings and emerging synchronies can occur during joint action, thus making multiple individuals act in similar ways without any intention to do so, which could be viewed as a case of emergent coordination. Other processes such as *interpersonal entrainment mechanisms* can lead to emergent coordination without requiring awareness: A famous example is the one of two people sitting in rocking chairs in the same room, who sometimes unconsciously synchronize their rocking frequency (Richardson et al., 2007); Another striking example is when two people walk side by side and sometimes unconsciously synchronize their steps (van Ulzen et al., 2008). Another source of unconscious emergent coordination which is worth mentioning here is the case of *perception-action matching* (Prinz, 1997; Jeannerod, 1999; Rizzolatti and Sinigaglia, 2010). It is a situation where actions performed by a first agent and observed by a second one are considered to be mentally matched onto the second agent's own action repertoire, through the involvement of *mirror neurons* and other mental processes that enable the induction of the same action tendencies in the two agents. All these processes are thought to make agents' behavior more similar and thus more predictable, which may facilitate joint action and coordination during action execution.

Humans thus have at their disposal a vast array of processes that they can use to promote interpersonal coordination. These processes range from automatic and unintentional on-the-fly alignments and synchronizations, to sophisticated forms of reasoning and advanced representational, conceptual and communicational skills. These processes are complementary and can be combined together to enable efficient joint action. Nevertheless, for human-robot interaction, this suggests that not all joint action situations may require some degree of awareness.

4.2. Human-Robot Joint Action

Human-robot joint action faces similar coordination challenges. We will explain now a way they can be translated to this case and quote some related implementation.

The robot needs to have the ability to represent itself and the human it interacts with. Doing so, it must be able to infer how each of these representations evolves along the joint action unfolding. The robot has to be able to consider perspective taking ability, knowing that representations evolve differently given each one point of view. Among others, Milliez et al. (2014) and Hiatt and Trafton (2010) endow a robot with the ability to construct a representation of other agents' mental states concerning the environment allowing it to pass the Sally and Anne test (Wimmer and Perner, 1983). Then, these mental states are used in Hiatt et al. (2011) to interpret and explain humans' behavior.

But the robot also needs to understand and take into account the effects of its own actions into the mental states of its partners, which involves a second-degree of awareness. This is done in Gray and Breazeal (2014) where the robot plays a competitive game with a human and chooses its action in order to manipulate the mental state of the human relative to the state of the world.

Each agent must also be able to asses the situation in terms of links with possible action: the objects that can be manipulated or moved, their location, the presence or absence of obstacles that could restrain some possibilities of movements. All these relate to the learned effects of actions presented in section 2 on affordances. In Sisbot et al. (2011) the robot uses the geometric information about the humans and the objects to construct symbolic knowledge as humans capabilities (an object is visible or reachable by someone), or relations between objects (an object is on/in another one). In Lemaignan et al. (2012) we have used this knowledge to anchor situated discourse during human-robot interaction. For example, if a human points at a mug saying "Give me that mug," the robot can understand that the human wants this mug and not another one. As a corollary, joint attention appears to be also key during human-robot joint action. This is because detecting a case of joint attention permits the robot to know that whatever information it can acquire within the joint attention space can be considered as also accessible to its interactor and thus as shared knowledge. Staudte and Crocker (2011) show that people interpret robot gaze in the same way as human gaze and that a congruent use of the robot gaze helps its human partner to understand spoken references. Mutlu et al. (2013) also show that the use of speech references in congruence with robot gaze enables to disambiguate spatial references in speech, and thus to improve task performance in joint action. They also put forward that robots in general might improve task performance and the quality of user experience during human-robot collaborative interaction by using action observation.

Another capacity needed by the robot, emphasized, among others, by Tomasello et al. (2005) as a prerequisite to joint action, is to be able to read its partner's actions. Gray et al. (2005) use the concept of mirror neurons, introduced by Gallese and Goldman (1998), to infer human action goals by matching the human movement to one of its own movements (even if the robot's morphology differs from that of the human). Hawkins et al. (2014) endow a robot with the capability to probabilistically infer what the human is doing and what he will do next in order to anticipate and prepare collaborative action. This capacity relates to probabilistic transitions learned within the type of world models used for robot decision-making in section 3. Again, this suggests that duplicating world models for each agent involved in the task (Lemaignan et al., 2012) can be a good strategy for human-robot joint action. This is in line with neuroscience proposals that a substantial component of awareness resides in the development of predictive models of agency for self and others (Seth et al., 2012), and in the ability to report about these states, predictions and decisions to self and to others (Shadlen and Kiani, 2011).

Complementarily, shared task representations are important. It means, if we paraphrase Knoblich et al. (2011)'s definition, that the robot should have access to some model of what each co-agents' respective task consist in and some abilities to monitor and predict each co-agent's actions with respect to the shared goal. (Nikolaidis and Shah, 2012) present a method allowing the robot to build a shared mental model of how to perform a collaborative task by looking at human performing the task and then use it when performing the task with a human.

We have seen that both in human-human and human-robot joint attention there are similar coordination constraints that apply. However, it appears that these constraints do not necessarily apply with the same strength. For instance, when two humans interact, they both know that they share some background knowledge such as cultural information, cultural knowledge, conventions, etc. Thus they can make assumptions from both sides on what the other knows or not. In contrast, it seems much more complicated to make similar assumptions in the human-robot interaction case.

Nevertheless, we have seen that human-human joint action sometimes involves planned joint action with explicit shared goals, action plans and attentions, and sometimes involve automatic synchronization or alignment processes between partners at a more sensory-motor level. Thus one might reasonably postulate that the integration of different types of learning and decision-making within robot cognitive architectures which has previously been applied to individual robotic tasks—such as the navigation task presented in section 3 or sequential decision-making tasks in Renaudo et al. (2014)—may be relevant in the context of human-robot interaction. This could enable the robot to automatically switch between automatic/habitual behavior and planned action depending on the requirement of the task, and thus display more behavioral flexibility and efficiency during joint action with humans.

Section 3 has put forward the hypothesis that the same coordination mechanisms for model-based and model-free reinforcement learning within robot architecture could be relevant both for non-social and social tasks in the context of the human-robot interaction task proposed by Alami (2013) and Lemaignan et al. (2017). Nevertheless, in this previous section the robot only achieved individual action plans, not joint action plans.

A more general illustration of awareness of each agent's actions' task-dependent effects and abilities that is required for joint action plans is shown in **Figure 8**. Again here, human and robot have to cooperate by putting some objects in certain placements where some are accessible only to the human or the robot. The robot has to elaborate a representation of different sub-spaces on the table so that it understands that some objects or places are accessible to the human. The robot tries to estimate visibility and reachability of the human and of itself (Pandey et al., 2013; Pandey and Alami, 2014) in order to determine the right places to use and where they can exchange objects. Also, the robot here has the capability to estimate the effort of the human in order to select the most pertinent places.

However, there is still a gap between such representations and those are required for the execution of an effective shared action plan. Indeed the robot should be able not only to compute the perspective of its human partner and use it to estimate how he can assess the current situation but also to estimate his current knowledge of the state of the task and the corresponding shared plan.

In Devin and Alami (2016) and Devin et al. (2017) we have developed, within the architecture described in Lemaignan et al. (2017), a framework that permits a robot to estimate the mental state of its human partner with respect to a given collaborative task achievement. We have moreover proposed a form of mental states which contains several task-relevant informations such as the states of the world, of the goals, actions and plans. To do so, the robot has to estimate and to permanently update the spatial perspective of its partners. It moreover has to constantly track their activity. Once these mental states representations are constructed and handled by the robot, it can use them to perform joint actions with humans. In the context of the present project, we have mostly investigated this in cases of collaborative objects manipulation. An advantage of the approach is to permit the robot to adapt online to the human's behavior and intention changes, while at the same time informing the human when needed in a non-intrusive manner, for instance by avoiding to give unnecessary information that the human could infer himself through observation or through deduction from past events.

As an illustration, let us consider a PR2 robot sharing with a human the goal of cleaning a table, that is, to first remove all objects on the table, then to sweep it, and afterwards to replace all objects back on the table. **Figure 9** shows the initial state of the

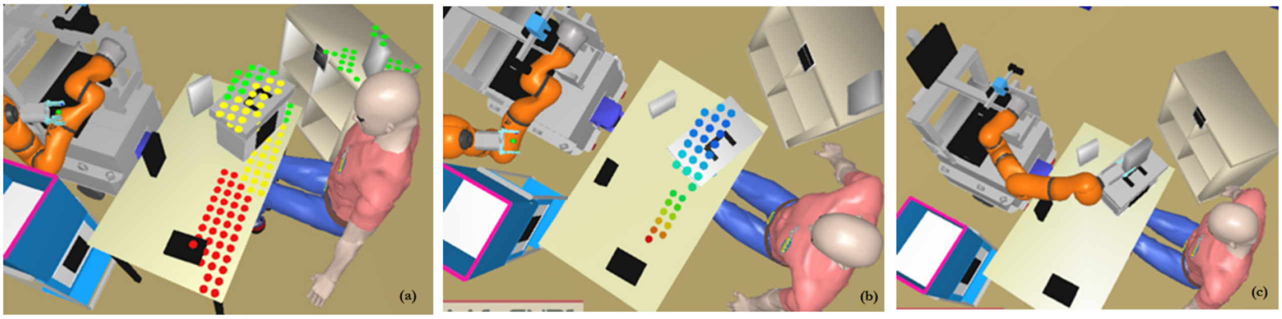


FIGURE 8 | Task of making an object accessible by the human to the robot (Pandey et al., 2013): **(a)** Places on the support planes where the human can put something with least effort. **(b)** Weighted points where the robot can support the human by taking the object. **(c)** The planner found a possible placement of the object on the box from where it is feasible for the robot to take. Note that, because of the object-closeness based weight assignment, this placement also reduces the human's effort to carry the object.

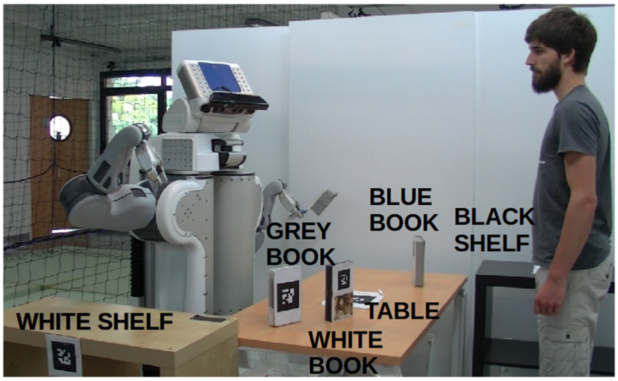


FIGURE 9 | Initial state of the world in the Clean the table scenario. In this task, the robot and the human share the goal of cleaning the table together.

world. On the table there is a blue book which is only reachable by the human, a gray book accessible only by the robot, and a white book reachable by both. Two actions are available to the robot: *pick-and-place* and *sweep*. The former can be executed by the robot only when the considered object and support on which to place the object are reachable by the robot. The latter can be executed on a surface only when it is again reachable by the robot and when there are not any objects on it. **Figure 10** illustrates the initial plan produced by the robot to achieve the goal.

The robot, equipped with such enlarged awareness ability, is not able to perform joint tasks more fluently. to reduce unnecessary communication and to choose the most pertinent way to inform about the state of the plan, to produce a less intrusive behavior of the robot but also potentially detect situations where human lacks an information allowing him to act and also the robot can in certain cases prevent human mistakes due to a wrong evaluation of the current state of the task.

These contributions involved pre-defined world, task and human models so that the robot can plan complex action plans involving collaborative human-robot task achievement with a human-aware task planner (HATP) (Alami et al., 2011; Lallement et al., 2014) and the associated high-level robot controller (Devin and Alami, 2016; Devin et al., 2017). This however did not involve a learning process. We have proposed in section 3 an extension of this work by considering that the subparts of the action sequence that are repeatedly performed by the robot in the same manner in this condition can be learned by the model-free habit learning system of their architecture. This is similar to habits learned by humans in conditions where repetitive behaviors are always occurring in the same context and in the same manner. This could enable the robot to autonomously detect and thus be aware of which situations are stable enough and repetitive enough to avoid systematically using the slow and costly action planning system. In addition, this framework should also enable the robot to automatically detect when environmental changes require to break the habit and switch back to the planning of new action sequences. Nevertheless, an extension of this work which is still under investigation consist in extending this framework by also enabling the robot to represent a world model associated to the human's actions' effects. This should permit to use model-based reinforcement learning to refine the world, the task and the human models used by HATP and the robot supervision system in order to find other action plans that could not be anticipated by the human experimenter. This could also lead to further awareness by the robot of which joint action plans are predictable by the human, and which should appear as new.

5. SELF-AWARE DECISION MAKING

5.1. General Approach

Planning in the field of AI is usually considered as the problem of building a sequence of actions selected from a predefined set in order to achieve a goal specified by a user or an external system (see Khamassi et al., 2016; Ingrand and Ghallab, 2017 for recent reviews). Classical planning is mostly based on First-order Predicate Logic or extensions thereof. If there are uncertainties on states, or on action outcomes, a probabilistic formulation is used and MDPs/POMDPs are the main tools.

The question addressed here is how can a system decide for its *own* goals, without being requested by an external agent?

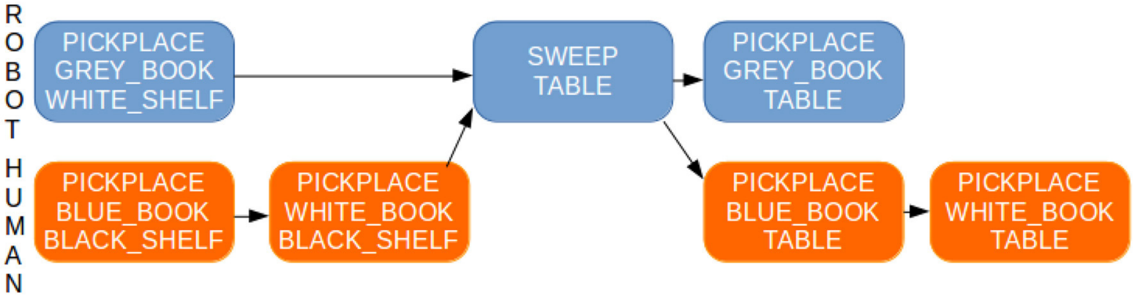

ROBOT
HUMAN
PICKPLACE GREY_BOOK WHITE_SHELF
SWEEP TABLE
PICKPLACE GREY_BOOK TABLE
PICKPLACE BLUE_BOOK BLACK_SHELF
PICKPLACE WHITE_BOOK BLACK_SHELF
PICKPLACE BLUE_BOOK TABLE
PICKPLACE WHITE_BOOK TABLE

FIGURE 10 | Shared plan computed by the robot to solve the joint goal: first removing the three objects (books) that are located on the table, then sweeping the table in order to clean it and finally placing the objects back on the table. While cooperatively achieving the task, the robot will be able to detect and assess correctly why the human partner stays idle, for instance in cases where, due to a momentary absence, the human may have missed the fact that robot has swept the table.

How can it decide to change goals dynamically? These questions are important because their answers determine if the agent is capable of a form of volition. Addressing them has lead to design a system capable of meta-reasoning to reflect on its objectives and on the way it is accomplishing them. In other words, the system described next is reasoning on its own motivations and actions, a feature we believe is related to self-awareness.

We want to build a system able to reach potentially concurrent goals and to manage resources such as energy and time, in an uncertain dynamic world. We aim for autonomous initiative and decision-making, so that the agent does not only react to particular stimuli or direct external requests, but most importantly selects by itself goals to achieve.

We consider the notion of *motivation* as the basis for bootstrapping the system's behavior, the trigger for a capacity of taking initiatives. The question of internal motivations has often been overlooked in the autonomous robotics literature: motivations are usually identified as simple drives emerging from external stimuli, whose dynamics are entirely dictated by the metabolism (e.g., decreasing energy level) and the occasional unconditional rewarding signals issued from the environment (such as locations for energy charging). The resulting systems are thus not purely reactive, but they can neither be considered as deliberative and motivationally autonomous because they lack an evaluation and selection among motivations. The selection is rather usually based on inhibitory signals resulting from external stimuli, such as in the multiple implementations of the subsumption architecture (Brooks, 1986).

Here, we want to investigate the potential advantage that an artificial system could have in developing its own preferences, i.e., to associate virtual rewards (to be distinguished from reward predictions used in actor-critic models, for example) to specific states which seem to have a key role in obtaining long-term rewards and should thus become intrinsically rewarding. These virtual rewards would be created by the motivational system, while the learning systems would remain unaware of the real or virtual aspect of the rewards they are manipulating. A possible advantage could be to set key-points where a reset of the reward discount mechanisms would be made, thus avoiding the problem of the discounted reward vanishing when trying to learn to reach very long-term goals.

This could account for example for the behavior of rats in the task studied in , where the stimulus seems to become a reward in itself, even when the food is not consumed. These virtual rewards could then be used for learning by model-based and model-free systems, as has been proposed by Lesaint et al. (2014) to account for these rats' behavior in a similar manner to the one presented in section 3. Virtual rewards could also explain how getting more money, a normally intermediate step which can indirectly lead to unconditional rewards like food, can become a reward in itself.

We focus here on the higher level of the robot cognitive architecture and propose to transform it into a deliberation system involving a self-awareness capacity. For this we hypothesize two layers of decision-making: (i) a higher level one called *deliberation* layer for solving multiple goal situations given motivations (using an "intentional module," context and long-term objectives, producing a "goal agenda" as input to (ii) a lower level goal-oriented planning system called the *operational* layer which will decide of the more precise course of actions to achieve the goals. This planning system is associated with a supervisory control system, which enables to control action execution as in classical systems.

The notion of *motivation* proposed in this paper, is a structure consisting of (individual or chained) goals, which may be permanently active or not, and to which we associate rewards. We aim to predict the precise effects of the resolution of a goal on the world and on other motivations, in order to compute a high-level plan, employing goal-reaching *policies* in the same way that we usually use *actions* in an MDP.

We hence develop an architecture that:

1. handles motivations,
2. computes possible policies for each motivation,
3. predicts the behavior of each policy and its effect on motivations,
4. predicts the effects of a chain of policies,
5. finds an optimal arrangement of these policies, maximizing the sum of the rewards obtained by the related motivations for a given time-horizon.

5.2. Motivations

Motivations are modeled as finite state machines corresponding to specific objectives, which can be permanent and basic, such as "maintain a high battery level," or complex and chained, such as "activate device A and then device B," etc.

The state of a motivation changes when there is a relevant change in the state of the world. To check if the conditions are met for changing the motivation state, an observation of world state transitions (ws, a, ws') is required, where ws is the initial world state, a is the executed action, and ws' is the resulting world state. World state transitions provide information that can trigger *motivation transitions*, i.e., changes in motivation states from ms to ms'. A *motivation transition* can be defined as an expression: $(ms, (ws, a, ws')) \rightarrow ms'$. We associate a positive or negative *reward* r to each motivation transition, reflecting its importance.

A rewarded motivation transition rmt starting from $ms1$ and leading to $ms2$ is called an *available-rmt* when the current motivation state is $ms1$. It *becomes activated* (or *triggered*) when the corresponding world transition (ws, a, ws') happens, changing the motivation state to $ms2$ and obtaining the corresponding positive or negative reward (r). The maximization of the sum of these rewards will be sought by the deliberation system.

5.3. Decision System

The architecture of the decision-making system is organized into the following modules (**Figure 11**):

- An *intentional module*, which manages the agent's objectives in the form of motivations. It is embedded in the deliberation modules (see next). It creates a list of motivations msv, containing the current states of all motivations. Consequently, given a msv, it is possible to know all the active rewarded motivation transitions originating from those current states, called *available-rmts*. This module is also responsible for keeping motivations up-to-date, depending on the world state evolution.
- An *operational module*, which computes policies based on the motivations automata, and computes predictions on resulting policies. It's based on an MDP.
- A *deliberation module*. Its role is to provide to the operational module rewarded word transitions rwt to reach, to enable it to build predictable solutions that will trigger the corresponding rewarded motivation transitions rmt. The deliberation module then computes the effects of these policies on the world state and on all motivations. These policies are used as macro-actions to compute a conditional high-level plan for maximizing the sum of the motivation rewards. This plan is called *policy agenda* handed to the supervisory system for execution. Thus, this module actually reasons on the active motivations, and on the best way to satisfy them using the policies the operation module can offer to achieve them. In other words, this is a meta-reasoning capacity, which we believe a core feature of self-awareness. The robot is not simply driven by its motivations, neither by a classical planning ability which determines a course of actions to achieve a goal. The robot determines its own goals by pondering how to satisfy its motivations and based on its planning results.

In summary, the actions are based on motivations that are driving the system's decisions. Motivations trigger the computation of policies to achieve them. Deliberation evaluates the policies to select the more rewarding actions. This achieves a meta reasoning capacity.

6. COGNITIVE ARCHITECTURE

The RoboErgoSum project employs a cognitive architecture designed for providing a robot with the necessary skills for autonomous activity in an unknown environment. The software architecture of the project is shown in **Figure 12**. Although we present an architecture unifying the modules detailed in the previous sections, a validation of the global architecture is yet to be done. Nevertheless, parts of this architecture were validated separately, as detailed at the end of this section.

The architecture contains modules for:

- sensing and acting in the environment (Sensorial perception and Motor modules),
- sensorimotor learning (sensorimotor learning module),
- symbolic knowledge generation and management (blue modules: Spatial reasoning and knowledge, Knowledge base)
- decision and action planning (green modules: Human-aware task planning, Reinforcement Learning model-free decision making system, Human-aware motion and manipulation planning),
- controlling the modules (Supervision system),
- goal management (Motivation module),
- dialogue management.

The interconnections between the modules are structured as follows.

The **Sensorial perception module** contains the innate set of perceptual abilities for perceiving the environment (visual perception and proprioception). The **Motor module** contains the innate set of action primitives available to the robot, which allow it to interact with the environment.

The **Sensorimotor learning** module processes the available pre-processed inputs (i.e., objects detected, actions performed, measured effects) to discover and learn which interactions are available to the robot in the current environment (i.e., affordance learning). It also generates the set of available actions that were learned after the interaction with the environment, together with their pre-conditions and post-conditions.

The **Spatial reasoning and knowledge** and the **Knowledge base** modules (Lemaignan et al., 2017) generate and store symbolic data about the perceived environment. This data is then used in the action planning phase by the corresponding modules: **Human-aware task planning module** (Lallement et al., 2014), and the **Human-aware motion and manipulation planning module** (Sisbot et al., 2007; Sisbot and Alami, 2012). Knowledge about the current state and the available actions is used by the **Reinforcement Learning model-free decision making system**.

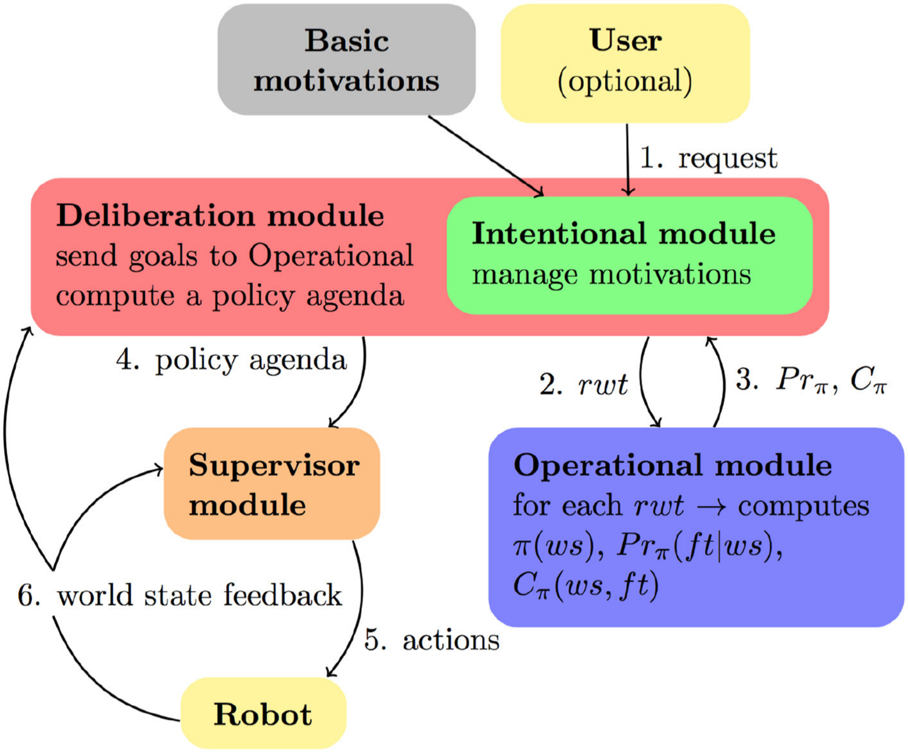


FIGURE 11 | Decision-making architecture including operational, intentional and deliberation modules. The deliberation module implements a meta-reasoning capability.

The **Supervision system** communicates with the aforementioned modules to decide which action planning system to employ, to perform on-line plan correction, and to monitor the activity of humans with which it interacts.

The **Motivation** module manages the set of goals that have to be achieved by the robot. Together with the action planning modules, it computes the optimal set of actions to perform, so as to obtain the highest reward in the given time horizon.

We validated several pieces of this architecture, using different sets of modules. We employed in an affordance-learning context the combination of modules responsible for Sensorial perception, Motor action execution, and Sensorimotor learning (the 3 yellow modules on the top of the **Figure 12**), previously described in section 2 (Chavez-Garcia et al., 2016b,a). Similarly, in a human-robot collaboration setting, we employed the modules for Sensorial perception, Motor action execution, Spatial reasoning and knowledge, Knowledge base, Supervision system, Human-aware task planning, Human-aware motion and manipulation planning, Motivation, and Dialogue Manager (Alami, 2013; Devin and Alami, 2016; Lemaignan et al., 2017). We also linked these modules with a Reinforcement-Learning model-free decision making system (Renaudo et al., 2015a), as described in section 3.

In spite of these advancements, a validation of the global architecture remains to be done. This would require a considerable engineering effort for integrating the presented modules, as not all the interfaces between them are present today.

7. LESSONS LEARNED AND CONCLUSION

Affordance learning mechanisms presented in section 2 to learn effects of actions constitute a first level of awareness of the distinction between self, other agents, movable objects and fixed elements of the environment. The learned action effects can moreover be used as transition estimates between states of the environment which can be used as world models for other learning and decision-making components of the robot cognitive architecture.

A second level of awareness can be permitted by having the agent monitor various dynamic signals about the environment and its performance to decide which learning strategy is relevant at any given moment, between the model-based and model-free strategies presented in section 3. This not only provides more behavioral flexibility and decisional autonomy, as we have previously argued (Renaudo et al., 2014; Khamassi et al., 2016), but as we proposed here can constitute a way for the robot to further evaluate and report about how it learned a task, which strategies were efficient in particular circumstances. Further investigations in this direction should study whether this enable more generalization for the robot when it can recognize similar circumstances (measured through the same performance and task monitoring signals) in which it could attempt similar learning strategies successfully. A further progress in integration could permit these monitoring mechanisms to inform in return the affordance learning module

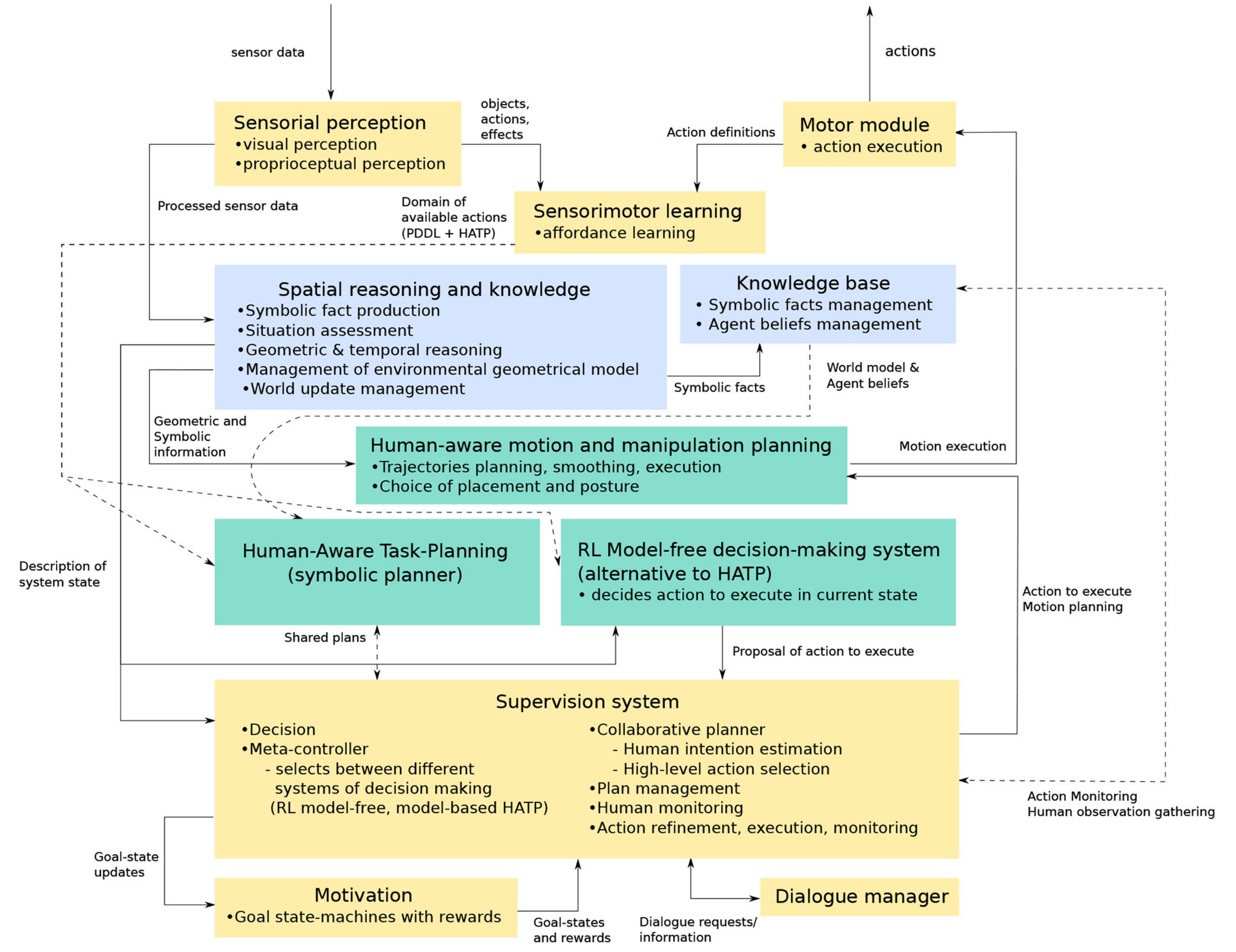

sensor data
actions
Sensorial perception
•visual perception
•proprioceptual perception
objects, actions, effects
Action definitions
Motor module
• action execution
Processed sensor data
Domain of available actions (PDDL + HATP)
Sensorimotor learning
•affordance learning
Spatial reasoning and knowledge
•Symbolic fact production
•Situation assessment
•Geometric & temporal reasoning
•Management of environmental geometrical model
•World update management
Knowledge base
• Symbolic facts management
• Agent beliefs management
Symbolic facts
World model & Agent beliefs
Geometric and Symbolic information
Human-aware motion and manipulation planning
•Trajectories planning, smoothing, execution
•Choice of placement and posture
Motion execution
Description of system state
Human-Aware Task-Planning (symbolic planner)
RL Model-free decision-making system (alternative to HATP)
• decides action to execute in current state
Action to execute Motion planning
Shared plans
Proposal of action to execute
Supervision system
•Decision
•Meta-controller
- selects between different systems of decision making (RL model-free, model-based HATP)
•Collaborative planner
- Human intention estimation
- High-level action selection
•Plan management
•Human monitoring
•Action refinement, execution, monitoring
Action Monitoring Human observation gathering
Goal-state updates
Motivation
•Goal state-machines with rewards
Goal-states and rewards
Dialogue requests/ information
Dialogue manager

FIGURE 12 | The global cognitive architecture employed in the RoboErgoSum project. Blue modules are responsible for generating and managing the symbolic knowledge. Decision-making modules are shown in green. Solid and dashed lines are used only to improve diagram readability where lines cross, and are otherwise identical in meaning.

to enrich the list of effects associated to actions with long-term effects in terms of different task resolutions. While this is still an ongoing part of the present project and requires further exploration, we argue here that such an integration of robot cognitive abilities should permit wider and long-term-oriented awareness of the agent to mentally represent what the tasks it can and cannot do with regards to its current capacities and past experience.

Besides, a particularly interesting lesson that we have learned from studying robot learning mechanisms in social and non-social tasks (section 4) is the observation that similar coordinations mechanisms of model-based and model-free learning strategies with a meta-controller can be relevant in both contexts. As the review of the human-human joint action literature suggests, joint action also involves both conscious model-based joint intention and unconscious action synchronization. Both are nevertheless important to enable intentional and unconscious signaling which enable each agent to be more predictable (and thus *readable*) by her coactor for efficient joint action. Application of the coordination of model-based and model-free learning mechanisms to human-robot interaction that we have initiated suggests that it could also permit the robot to become aware of which tasks performed in interaction with the human can be performed habitually, and which require a constant monitoring and reevaluation of possible action consequences through learned world models. This can further promote the development of internal models of what the human can and cannot do, which objects it can or cannot reach, as well as models of what the respective tasks of each of the co-agents.

Interestingly, some of the previous human-robot joint action experiments that we have previously done and summarized here suggest that a simple duplication of the robot's individual learning mechanisms presented here could be done within the robot's internal representations for each agent involved in the task. In other words, world models can be learned and generalized for each agent involved in the task (Lemaignan et al., 2012) in order to endow a robot with the capability to probabilistically infer what the human is doing and what he will do next in order to anticipate and prepare collaborative action. This is in line with neuroscience proposals that a substantial component of awareness resides in the development of predictive models of agency for self and others (Seth et al., 2012), and in the ability to report about these states, predictions and decisions to self and to others (Shadlen and Kiani, 2011).

Section 5 presented progress in the development of further awareness abilities, this time about the agent's decisions on its goals and motivations represented by finite state machines. We presented a system for managing multiple concurrent and permanent objectives, performing probabilistic reasoning with MDPs and capable of reasoning its plans to decide for the most rewarding actions. The deliberative system has a modular architecture, which separates the planning from the goal-managing entity, allowing for an easy integration into an existing robotic cognitive architecture.

Finally, we presented a global cognitive architecture designed to permit the integration of these different cognitive functions.

The whole work reported in this paper provides insights about how to achieve a self-aware system and to decipher what is awareness and what it is not, by monitoring the processes of the robot and recognizing when they solved a task with explicit deliberation and model-based strategies, or through unconscious model-free learning. It would be interesting to be able to measure the amount of integrated information in the robot cognitive architecture during these different processes, and see whether we can differentiate the two quantitatively, in agreement with the integrated information theory of Oizumi et al. (2014). This would need to be the subject of future research projects.

AUTHOR CONTRIBUTIONS

RA, RC, AC, BG, and MK designed the research. RC is the project leader. MA, R-OC-G, SD, RG, PL-V, and ER made developments and performed the experiments. All authors contributed to data analyses and interpretation. RA, MA, RC, R-OC-G, AC, SD, BG, RG, MK, PL-V, and ER wrote the manuscript.

FUNDING

This work has been funded by two DGA (French National Defence Armaments Procurement Agency) scholarships (PL-V, ER), by the French Agence Nationale de la Recherche (ANR–12–CORD–0030 Roboergosum Project and ANR–11–IDEX–000–02 Sorbonne–Universities SU–15–R–PERSU–14 Robot Parallearning Project), by Labex SMART (ANR–11–LABX–65), and by the Centre National de la Recherche Scientifique (Mission pour l'Interdisciplinarité) ROBAUTISTE Project.

ACKNOWLEDGMENTS

The authors would like to thank Lucie Charles, Elisabeth Pacherie, Claire Sergent, and Catherine Tallon-Baudry for contributions and useful discussions.

REFERENCES

Alami, R. (2013). "On human models for collaborative robots," in *2013 International Conference on Collaboration Technologies and Systems, CTS 2013, May 20-24, 2013* (San Diego, CA), 191–194. doi: 10.1109/CTS.2013.6567228

Alami, R., Warnier, M., Guitton, J., Lemaignan, S., and Sisbot, E. A. (2011). "When the robot considers the human...," in *Proceedings of the 15th International Symposium on Robotics Research* (Flagstaff, AZ).

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Andries, M., Chavez-Garcia, R. O., Chatila, R., Giusti, A., and Gambardella, L. M. (2018). Affordance equivalences in robotics: a formalism. *Front. Neurorobot.* 12:26. doi: 10.3389/fnbot.2018.00026

Balleine, B. W., and O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35:48. doi: 10.1038/npp.2009.131

Bratman, M. E. (2014). *Shared Agency: A Planning Theory of Acting Together*. New York, NY: Oxford University Press.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106, 1465–1477. doi: 10.1016/j.cognition.2007.05.012

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE J. Robot. Automat.* 2, 14–23. doi: 10.1109/JRA.1986.1087032

Caluwaerts, K., Staffa, M., N'Guyen, S., Grand, C., Dollé, L., Favre-Félix, A., et al. (2012). A biologically inspired meta-control navigation system for the psikharpax rat robot. *Bioinspir. Biomimet.* 7:025009. doi: 10.1088/1748-3182/7/2/025009

Chavez-Garcia, R. O., Andries, M., Luce-Vayrac, P., and Chatila, R. (2016a). "Discovering and manipulating affordances," in *International Symposium on Experimental Robotics (ISER)* (Tokyo).

Chavez-Garcia, R. O., Luce-Vayrac, P., and Chatila, R. (2016b). "Discovering affordances through perception and manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Daejeon).

Chella, A., and Manzotti, R. (eds.). (2007). *Artificial Consciousness*. Exeter: Imprint Academic.

Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.

Clodic, A., Pacherie, E., Alami, R., and Chatila, R. (2017). *Key Elements for Human-Robot Joint Action*. Cham: Springer International Publishing.

Comaniciu, D., Meer, P., and Member, S. (2002). Mean shift: a robust approach toward feature space analysis. *Patt. Anal. Mach. Intell. IEEE Trans.* 24, 603–619. doi: 10.1109/34.1000236

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711. doi: 10.1038/nn1560

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Netw.* 22, 213–219. doi: 10.1016/j.neunet.2009.03.004

Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1–37. doi: 10.1016/S0010-0277(00)00123-2

Devin, S., and Alami, R. (2016). "An implemented theory of mind to improve human-robot shared plans execution," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, March 7-10, 2016* (Christchurch), 319–326.

Devin, S., Clodic, A., and Alami, R. (2017). "About decisions during human-robot shared plan achievement: who should act and how?," in *Social Robotics - 9th International Conference, ICSR 2017, November 22-24, 2017, Proceedings* (Tsukuba), 453–463.

Dollé, L., Chavarriaga, R., Guillot, A., and Khamassi, M. (2018). Interactions of spatial strategies producing generalization gradient and blocking: a computational approach. *PLoS Comput. Biol.* 14:e1006092. doi: 10.1371/journal.pcbi.1006092

Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., and Guillot, A. (2010). Path planning versus cue responding: a bioinspired model of switching between navigation strategies. *Biol. Cybern.* 103, 299–317. doi: 10.1007/s00422-010-0400-z

Fikes, R. E., and Nilsson, N. J. (1971). "Strips: a new approach to the application of theorem proving to problem solving," in *Proceedings of the 2Nd International Joint Conference on Artificial Intelligence, IJCAI'71* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 608–620.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., and Willuhn, I. (2011). A selective role for dopamine in stimulus-reward learning. *Nature* 469, 53–57. doi: 10.1038/nature09588

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* 2, 493–501. doi: 10.1016/S1364-6613(98)01262-5

Gibson, J. (1977). "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, eds R. E. Shaw and J. Bransford (Hillsdale, NJ: Lawrence Erlbaum Associates), 67–82.

Gray, J., and Breazeal, C. (2014). Manipulating mental states through physical action. *Int. J. Soc. Robot.* 6, 315–327. doi: 10.1007/s12369-014-0234-2

Gray, J., Breazeal, C., Berlin, M., Brooks, A., and Lieberman, J. (2005). "Action parsing and goal inference using self as simulator," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on* (Nashville, TN: IEEE), 202–209.

Grisetti, G., Stachniss, C., and Burgard, W. (2007). Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Trans. Rob*. 23, 34–46. doi: 10.1109/TRO.2006.889486

Grosz, B. J., and Kraus, S. (1996). Collaborative plans for complex group action. *Artif. Intell.* 86, 269–357. doi: 10.1016/0004-3702(95)00103-4

Haggard, P., and Tsakiris, M. (2009). The experience of agency: feelings, judgments, and responsibility. *Curr. Direct. Psychol. Sci.* 18, 242–246. doi: 10.1111/j.1467-8721.2009.01644.x

Hawkins, K. P., Bansal, S., Vo, N. N., and Bobick, A. F. (2014). "Anticipating human actions for collaboration in the presence of task and sensor uncertainty," in *Robotics and automation (ICRA), 2014 ieee international conference on* (Hong Kong: IEEE), 2215–2222.

Hiatt, L. M., Harrison, A. M., and Trafton, J. G. (2011). "Accommodating human variability in human-robot teams through theory of mind," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Vol. 22 (Barcelona), 2066.

Hiatt, L. M., and Trafton, J. G. (2010). "A cognitive model of theory of mind," in *Proceedings of the 10th International Conference on Cognitive Modeling* (Philadelphia, PA: Drexel University), 91–96.

Hood, B. (2012). *The Self Illusion: How the Social Brain Creates Identity*. Oxford, UK: Oxford University Press.

Ingrand, F., and Ghallab, M. (2017). Deliberation for autonomous robots: a survey. *Artif. Intell.* 247, 10–44. doi: 10.1016/j.artint.2014.11.003

Jeannerod, M. (1999). The 25th bartlett lecture. *Q. J. Exp. Psychol. Sect. A* 52, 1–29. doi: 10.1080/713755803

Khamassi, M., Girard, B., Clodic, A., Sandra, D., Renaudo, E., Pacherie, E., et al. (2016). Integration of action, joint action and learning in robot cognitive architectures. *Intell. Assoc. Pour Recher. Sci. Cogn.* 2016, 169–203. Available online at: http://intellectica.org/fr/integration-de-l-action-de-l-action-conjointe

Knoblich, G., Butterfill, S., and Sebanz, N. (2011). Psychological research on joint action: theory and data. *Psychol. Learn. Motivat. Adv. Res. Theory* 54:59. doi: 10.1016/B978-0-12-385527-5.00003-6

Kober, J., Bagnell, D., and Peters, J. (2013). Reinforcement learning in robotics: a survey. *IJRR J.* 11, 1238–1274. doi: 10.1177/0278364913495721

Koch, C., Massimini, M., Boly, M., and G., T. (2016). Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* 17, 307–321. doi: 10.1038/nrn.2016.22

Lallement, R., de Silva, L., and Alami, R. (2014). HATP: an HTN planner for robotics. *CoRR* abs/1405.5345.

Lehman, J. F., Laird, J., and Rosenbloom, P. (2006). *A Gentle Introduction to Soar : An Architecture for Human Cognition: 2006 Update*. Available online at: http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/GentleIntroduction-2006.pdf

Lemaignan, S., Ros, R., Sisbot, E. A., Alami, R., and Beetz, M. (2012). Grounding the interaction: anchoring situated discourse in everyday human-robot interaction. *Int. J. Soc. Robot.* 4, 181–199. doi: 10.1007/s12369-011-0123-x

Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., and Alami, R. (2017). Artificial cognition for social human-robot interaction: an implementation. *Artif. Intell.* 247, 45–69. doi: 10.1016/j.artint.2016.07.002

Lesaint, F., Sigaud, O., Flagel, S. B., Robinson, T. E., and Khamassi, M. (2014). Modelling individual differences in the form of pavlovian conditioned approach responses: a dual learning systems approach with factored representations. *PLoS Comput. Biol.* 10:e1003466. doi: 10.1371/journal.pcbi.1003466

Lewis, P. R., Chandra, A., Parsons, S., Robinson, E., Glette, K., Bahsoon, R., et al. (2011). "A survey of self-awareness and its application in computing systems," in *Self-Adaptive and Self-Organizing Systems Workshops (SASOW), 2011 Fifth IEEE Conference on* (IEEE), 102–107.

Marks-Tarlow, T. (1999). The self as a dynamical system. *Nonlin. Dyn. Psychol. Life Sci.* 3, 311–345. doi: 10.1023/A:1021958829905

Milliez, G., Warnier, M., Clodic, A., and Alami, R. (2014). "A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on* (Edinburgh: IEEE), 1103–1109.

Morin, A. (2006). Levels of consciousness and self-awareness: a comparison and integration of various neurocognitive views. *Conscious. Cogn.* 15, 358–371. doi: 10.1016/j.concog.2005.09.006

Mutlu, B., Terrell, A., and Huang, C.-M. (2013). "Coordination mechanisms in human-robot collaboration," in *Proceedings of the Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction* (Tokyo), 1–6.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Nikolaidis, S., and Shah, J. (2012). "Human-robot teaming using shared mental models," in *IEEE/ACM International Conference on Human-Robot Interaction, Workshop on Human-Agent-Robot Teamwork* (Boston, MA).

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol.* 10:e1003588. doi: 10.1371/journal.pcbi.1003588

Pandey, A. K., and Alami, R. (2014). Towards human-level semantics understanding of human-centered object manipulation tasks for HRI: reasoning about effect, ability, effort and perspective taking. *I. J. Soc. Robot.* 6, 593–620. doi: 10.1007/s12369-014-0246-y

Pandey, A. K., Ali, M., and Alami, R. (2013). Towards a task-aware proactive sociable robot based on multi-state perspective-taking. *I. J. Soc. Robot.* 5, 215–236. doi: 10.1007/s12369-013-0181-3

Papon, J., Abramov, A., Schoeler, M., and Worgotter, F. (2013). "Voxel cloud connectivity segmentation - Supervoxels for point clouds," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 2027–2034.

Pezzulo, G., Donnarumma, F., and Dindo, H. (2013). Human sensorimotor communication: a theory of signaling in online social interactions. *PLoS ONE* 8:e79876. doi: 10.1371/journal.pone.0079876

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154. doi: 10.1080/713752551

Renaudo, E., Devin, S., Girard, B., Chatila, R., Alami, R., Khamassi, M., et al. (2015a). "Learning to interact with humans using goal-directed and habitual behaviors," in *RoMan 2015, Workshop on Learning for Human-Robot Collaboration* (Kobe).

Renaudo, E., Girard, B., Chatila, R., and Khamassi, M. (2014). "Design of a control architecture for habit learning in robots," in *Biomimetic and Biohybrid Systems, LNAI Proceedings* (Milan), 249–260. doi: 10.1007/978-3-319-09435-9_22

Renaudo, E., Girard, B., Chatila, R., and Khamassi, M. (2015b). "Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture," in *Biologically Inspired Cognitive Architectures BICA 2015*, (Lyon), 178–184.

Renaudo, E., Girard, B., Chatila, R., and Khamassi, M. (2015c). "Which criteria for autonomously shifting between goal-directed and habitual behaviors in robots?," in *5th International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB)* (Providence, RI), 254–260.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., and Schmidt, R. C. (2007). Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Hum. Movem. Sci.* 26, 867–891. doi: 10.1016/j.humov.2007.07.002

Rizzolatti, G., and Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nat. Rev. Neurosci.* 11, 264–274. doi: 10.1038/nrn2805

Sahin, E., Cakmak, M., Dogar, M. R., Ugur, E., and Ucoluk, 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

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Front. Psychol.* 2:395. doi: 10.3389/fpsyg.2011.00395

Shadlen, M. N., and Kiani, R. (2011). "Consciousness as a decision to engage," in *Characterizing Consciousness: From Cognition to the Clinic?*, eds S. Dehaene and Y. Christen (Berlin; Heidelberg: Springer), 27–46.

Shoda, Y., LeeTiernan, S., and Mischel, W. (2002). Personality as a dynamical system: emergence of stability and distinctiveness from intra and interpersonal interactions. *Pers. Soc. Psychol. Rev.* 6, 316–325. doi: 10.1207/S15327957PSPR0604_06

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). "Indoor segmentation and support inference from RGBD images," in *Proceedings of the 12th European Conference on Computer Vision* (Florence), 746–760.

Sisbot, E. A., and Alami, R. (2012). A human-aware manipulation planner. *IEEE Trans. Robot.* 28, 1045–1057. doi: 10.1109/TRO.2012.2196303

Sisbot, E. A., Marin-Urias, L. F., Alami, R., and Siméon, T. (2007). A human aware mobile robot motion planner. *IEEE Trans. Robot.* 23, 874–883. doi: 10.1109/TRO.2007.904911

Sisbot, E. A., Ros, R., and Alami, R. (2011). "Situation assessment for human-robot interactive object manipulation," in *RO-MAN, 2011 IEEE* (Atlanta, GA: IEEE), 15–20.

Staudte, M., and Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition* 120, 268–291. doi: 10.1016/j.cognition.2011.05.005

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, Vol. 1. Cambridge: MIT Press.

Tomasello, M., and Carpenter, M. (2007). Shared intentionality. *Dev. Sci.* 10, 121–125. doi: 10.1111/j.1467-7687.2007.00573.x

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). In search of the uniquely human. *Behav. Brain Sci.* 28, 721–727. doi: 10.1017/S0140525X05540123

Van Gulick, R. (2017). "Consciousness," in *The Stanford Encyclopedia of Philosophy*, *Summer 2017 Edn.*, ed E. N. Zalta (Metaphysics Research Lab, Stanford University). Available online at: https://plato.stanford.edu/entries/consciousness/

van Hoof, H., Kroemer, O., and Peters, J. (2014). Probabilistic segmentation and targeted exploration of objects in cluttered environments. *IEEE Trans. Robot.* 30, 1198–1209. doi: 10.1109/TRO.2014.2334912

van Ulzen, N. R., Lamoth, C. J., Daffertshofer, A., Semin, G. R., and Beek, P. J. (2008). Characteristics of instructed and uninstructed interpersonal coordination while walking side-by-side. *Neurosci. Lett.* 432, 88–93. doi: 10.1016/j.neulet.2007.11.070

Vesper, C. (2014). "How to support action prediction: evidence from human coordination tasks," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on* (Edinburgh: IEEE), 655–659.

Viejo, G., Khamassi, M., Brovelli, A., and Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Front. Behav. Neurosci.* 9:225. doi: 10.3389/fnbeh.2015.00225

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

Yin, H. H., and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.* 7, 464–476. doi: 10.1038/nrn1919

Conflict of Interest Statement: 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.

frontiers in Neurorobotics

METHODS
published: 30 August 2018
doi: 10.3389/fnbot.2018.00045

Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop

Martin Biehl[1]*, *Christian Guckelsberger*[2], *Christoph Salge*[3,4], *Simón C. Smith*[4,5] *and Daniel Polani*[4]

[1] Araya Inc., Tokyo, Japan, [2] Computational Creativity Group, Department of Computing, Goldsmiths, University of London, London, United Kingdom, [3] Game Innovation Lab, Department of Computer Science and Engineering, New York University, New York, NY, United States, [4] Sepia Lab, Adaptive Systems Research Group, Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom, [5] Institute of Perception, Action and Behaviour, School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom

Active inference is an ambitious theory that treats perception, inference, and action selection of autonomous agents under the heading of a single principle. It suggests biologically plausible explanations for many cognitive phenomena, including consciousness. In active inference, action selection is driven by an objective function that evaluates possible future actions with respect to current, inferred beliefs about the world. Active inference at its core is independent from extrinsic rewards, resulting in a high level of robustness across e.g., different environments or agent morphologies. In the literature, paradigms that share this independence have been summarized under the notion of intrinsic motivations. In general and in contrast to active inference, these models of motivation come without a commitment to particular inference and action selection mechanisms. In this article, we study if the inference and action selection machinery of active inference can also be used by alternatives to the originally included intrinsic motivation. The perception-action loop explicitly relates inference and action selection to the environment and agent memory, and is consequently used as foundation for our analysis. We reconstruct the active inference approach, locate the original formulation within, and show how alternative intrinsic motivations can be used while keeping many of the original features intact. Furthermore, we illustrate the connection to universal reinforcement learning by means of our formalism. Active inference research may profit from comparisons of the dynamics induced by alternative intrinsic motivations. Research on intrinsic motivations may profit from an additional way to implement intrinsically motivated agents that also share the biological plausibility of active inference.

Keywords: intrinsic motivation, free energy principle, active inference, predictive information, empowerment, perception-action loop, universal reinforcement learning, variational inference

OPEN ACCESS

Edited by:
Antonio Chella,
Universit degli Studi di Palermo, Italy

Reviewed by:
Karl Friston,
University College London, United Kingdom
Alessandro Di Nuovo,
Sheffield Hallam University, United Kingdom

****Correspondence:***
Martin Biehl
martin@araya.org

Received: *18 April 2018*
Accepted: *02 July 2018*
Published: *30 August 2018*

Citation:
Biehl M, Guckelsberger C, Salge C, Smith SC and Polani D (2018) Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop. Front. Neurorobot. 12:45. doi: 10.3389/fnbot.2018.00045

1. INTRODUCTION

Active inference (Friston et al., 2012), and a range of other formalisms usually referred to as intrinsic motivations (Storck et al., 1995; Klyubin et al., 2005; Ay et al., 2008), all aim to answer a similar question: "Under minimal assumptions, how should an agent act?" More practically, they relate to what would be a universal way to generate behaviour for an agent or robot that appropriately deals with its environment, i.e., acquires the information needed to act and acts toward an intrinsic goal. To this end, both the free energy principle and intrinsic motivations aim to bridge the gap between giving a biologically plausible explanation for how real organism deal with the problem and providing a formalism that can be implemented in artificial agents. Additionally, they share a range of properties, such as an independence of a priori semantics and being defined purely on the dynamics of the agent environment interaction, i.e., the agent's perception-action loop.

Despite these numerous similarities, as far as we know, there has not been any unified or comparative treatment of those approaches. We believe this is in part due to a lack of an appropriate unifying mathematical framework. To alleviate this, we present a technically complete and comprehensive treatment of active inference, including a decomposition of its perception and action selection modes. Such a decomposition allows us to relate active inference and the inherent motivational principle to other intrinsic motivation paradigms such as empowerment (Klyubin et al., 2005), predictive information (Ay et al., 2008), and knowledge seeking (Storck et al., 1995; Orseau et al., 2013). Furthermore, we are able to clarify the relation to universal reinforcement learning (Hutter, 2005). Our treatment is deliberately comprehensive and complete, aiming to be a reference for readers interested in the mathematical fundament.

A considerable number of articles have been published on active inference (e.g., Friston et al., 2012, 2015, 2016a,b, 2017a,b; Linson et al., 2018). Active inference defines a procedure for both perception and action of an agent interacting with a partially observable environment. The definition of the method, in contrast to other existing approaches (e.g., Hutter, 2005; Doshi-Velez et al., 2015; Leike, 2016), does not maintain a clear separation between the inference and the action selection mechanisms, and the objective function. Most approaches for perception and action selection are generally formed of three steps: The first step involves a learning or inference mechanism to update the agent's knowledge about the consequences of its actions. In a second step, these consequences are evaluated with respect to an agent-internal objective function. Finally, the action selection mechanism chooses an action depending on the preceding evaluation.

In active inference, these three elements are entangled. On one hand, there is the main feature of active inference: the combination of knowledge updating and action selection into a single mechanism. This single mechanism is the minimization of a "variational free energy" (Friston et al., 2015, p. 188). The "inference" part of the name is justified by the formal resemblance of the method to the variational free energy minimization (also known as evidence lower bound maximization) used in variational inference. Variational inference is a way to turn Bayesian inference into an optimization problem which gives rise to an approximate Bayesian inference method (Wainwright and Jordan, 2007). The "active" part is justified by the fact that the output of this minimization is a probability distribution over actions from which the actions of the agent are then sampled. Behaviour in active inference is thus the result of a variational inference-like process. On the other hand, the function (i.e., expected free energy) that induces the objective function in active inference is said to be "of the same form" as the variational free energy (Friston et al., 2017a, p. 2673) or even to "follow" from it (Friston et al., 2016b, p. 10). This suggests that expected free energy is the only objective function compatible with active inference.

In summary, perception and action in active inference intertwines four elements: variational approximation, inference, action selection, and an objective function. Besides these formal features, active inference is of particular interest for its claims on biological plausibility and its relationship to the thermodynamics of dissipative systems. According to Friston et al. (2012, Section 3) active inference is a "corollary" to the free energy principle. Therefore, it is claimed, actions must minimize variational free energy to resist the dispersion of states of self-organizing systems (see also Friston, 2013b; Allen and Friston, 2016). Active inference has also been used to reproduce a range of neural phenomena in the human brain (Friston et al., 2016b), and the overarching free energy principle has been proposed as a "unified brain theory" Friston (2010). Furthermore, the principle has been used in a hierarchical formulation as theoretical underpinning of the predictive processing framework (Clark, 2015, p. 305–306), successfully explaining a wide range of cognitive phenomena. Of particular interest for the present special issue, the representation of probabilities in the active inference framework is conjectured to be related to aspects of consciousness (Friston, 2013a; Linson et al., 2018).

These strong connections between active inference and biology, statistical physics, and consciousness research make the method particularly interesting for the design of artificial agents that can interact with- and learn about unknown environments. However, it is currently not clear to which extent active inference allows for modifications. We ask: how far do we have to commit to the precise combination of elements used in the literature, and what becomes interchangeable?

One target for modifications is the objective function. In situations where the environment does not provide a specific reward signal and the goal of the agent is not directly specified, researchers often choose the objective function from a range of *intrinsic motivations*. The concept of intrinsic motivation was introduced as a psychological concept by Ryan and Deci (2000), and is defined as "the doing of an activity for its inherent satisfactions rather than for some separable consequence." The concept helps us to understand one important aspect of consciousness: the assignment of affect to certain experiences, e.g., the experience of fun (Dennett, 1991) when playing a game. Computational approaches to intrinsic motivations (Oudeyer and Kaplan, 2009; Schmidhuber, 2010; Santucci et al., 2013) can be categorized roughly by the psychological motivations they are imitating, e.g., drives to manipulate and explore, the reduction of cognitive dissonance, the achievement of optimal incongruity, and finally motivations for effectance, personal causation, competence and self-determination. Intrinsic motivations have been used to enhance behaviour aimed at extrinsic rewards (Sutton and Barto, 1998), but their defining characteristic is that they can serve as a goal-independent motivational core for autonomous behaviour generation. This characteristic makes them good candidates for the role of value functions for the design of intelligent systems (Pfeifer et al., 2005). We attempt to clarify how to modify active inference to accommodate objective functions based on different intrinsic motivations. This may allow future studies to investigate whether and how altering the objective function affects the biological plausibility of active inference.

Another target for modification, originating more from a theoretical standpoint, is the variational formulation of active inference. As mentioned above, variational inference formulates Bayesian inference as an optimization problem; a family of probability distributions is optimized to approximate the direct, non-variational Bayesian solution. Active inference is formulated as an optimization problem as well. We consequently ask: is active inference the variational formulation of a direct (non-variational) Bayesian solution? Such a direct solution would allow a formally simple formulation of active inference without recourse to optimization or approximation methods, at the cost of sacrificing tractability in most scenarios.

To explore these questions, we take a step back from the established formalism, gradually extend the active inference framework, and comprehensively reconstruct the version presented in Friston et al. (2015). We disentangle the four components of approximation, inference, action selection, and objective functions that are interwoven in active inference.

One of our findings, from a formal point of view, is that expected free energy can be replaced by other intrinsic motivations. Our reconstruction of active inference then yields a unified formal framework that can accommodate:

- Direct, non-variational Bayesian inference in combination with standard action selection schemes known from reinforcement learning as well as objective functions induced by intrinsic motivations.
- Universal reinforcement learning through a special choice of the environment model and a small modification of the action selection scheme.
- Variational inference in place of the direct Bayesian approach.
- Active inference in combination with objective functions induced by intrinsic motivations.

We believe that our framework can benefit active inference research as a means to compare the dynamics induced by alternative action selection principles. Furthermore, it equips researchers on intrinsic motivations with additional ways for designing agents that share the biological plausibility of active inference.

Finally, this article contributes to the research topic: Consciousness in Humanoid Robots, in several ways. First, there have been numerous claims on how active inference relates to consciousness or related qualities, which we outlined earlier in the introduction. The most recent work by Linson et al. (2018), also part of this research topic, specifically discusses this relation, particularly in regards to assigning salience. Furthermore, intrinsic motivations (including the free energy principle for this argument) have a range of properties that relate to or are useful to a range of classical approaches recently summarized as as Good Old-Fashioned Artificial Consciousness (GOFAC, Manzotti and Chella, 2018). For example, embodied approaches still need some form of value-function or motivation (Pfeifer et al., 2005), and benefit from the fact that intrinsic motivations are usually universal yet sensitive in regards to an agent's embodiment. The enactive AI framework (Froese and Ziemke, 2009), another candidate for GOFAC, proposes further requirements on how value underlying motivation should be grounded in constitutive autonomy and adaptivity. Guckelsberger and Salge (2016) present tentative claims on how empowerment maximization relates to these requirements in biological systems, and how it could contribute to realizing them in artificial ones. Finally, the idea of using computational approaches for intrinsic motivation goes back to developmental robotics (Oudeyer et al., 2007), where it is suggested as way to produce a learning and adapting robot, which could offer another road to robot consciousness. Whether these Good Old-Fashioned approaches will ultimately be successful is an open question, and Manzotti and Chella (2018) asses them rather critically. However, extending active inference to alternative intrinsic motivations in a unified framework allows to combine features of these two approaches. For example it may bring together the neurobiological plausibility of active inference and the constitutive autonomy afforded by empowerment.

2. RELATED WORK

Our work is largely based on Friston et al. (2015) and we adopt the setup and models from it. This means many of our assumptions are due to the original paper. Recently, Buckley et al. (2017) have provided an overview of continuous-variable active inference with a focus on the mathematical aspects, rather than the relationship to thermodynamic free energy, biological interpretations or neural correlates. Our work here is in as similar spirit but focuses on the discrete formulation of active inference and how it can be decomposed. As we point out in the text, the case of direct Bayesian inference with separate action selection is strongly related to general reinforcement learning (Hutter, 2005; Leike, 2016; Aslanides et al., 2017). This approach also tackles unknown environments with- and in later versions also without externally specified reward in a Bayesian way. Other work focusing on unknown environments with rewards are e.g., (Ross and Pineau, 2008; Doshi-Velez et al., 2015). We would like to stress that we do not propose agents using Bayesian or variational inference as competitors to any of the existing methods. Instead, our goal is to provide an unbiased investigation of active inference with a particular focus on extending the inference methods, objective functions and action-selection mechanisms. Furthermore, these agents follow almost completely in a straightforward (if quite involved) way from the model in Friston et al. (2015). A small difference is the extension to parameterizations of environment and sensor dynamics. These parameterizations can be found in Friston et al. (2016b).

We note that work on planning as inference (Attias, 2003; Toussaint, 2009; Botvinick and Toussaint, 2012) is generally related to active inference. In this line of work the probability distribution over actions or action sequences that lead to a given goal specified as a sensor value is inferred. Since active inference also tries to obtain a probability distribution over actions the approaches are related. The formalization of the goal however differs, at least at first sight. How exactly the two approaches relate is beyond the scope of this publication.

3. STRUCTURE OF THIS ARTICLE

Going forward, we will first outline our mathematical notation in Section 4. We then introduce the perception-action loop, which contains both agent and environment in Section 5. In Section 6 we introduce the model used by Friston et al. (2015). We then show how to obtain beliefs about the consequences of actions via both (direct) Bayesian inference (Section 6.2) and (approximate) variational inference (Section 6.4). These beliefs are represented in the form of a set of complete posteriors. Such a set is a common object but usually does not play a prominent role in Bayesian inference. Here, it turns out to be a convenient structure for capturing the agent' knowledge and describing intrinsic motivations. Under certain assumptions that we discuss in Section 6.3 the direct Bayesian case specializes to the belief updating of the Bayesian universal reinforcement learning agent of Aslanides et al. (2017). We then discuss in Section 7 how those beliefs (i.e., the set of complete posteriors) can induce action-value functions (playing the role of objective functions) via a given intrinsic motivation function. We present standard (i.e., non-active inference) ways to select actions based on such action-value functions. Then we look at different instances of intrinsic motivation functions. The first is the "expected free energy" of active inference. For this we explicitly show how our formalism produces the original expression in Friston et al. (2015). Looking at the formulations of other intrinsic motivations it becomes clear that the expected free energy relies on expressions quite similar or identical to those that occur in other intrinsic motivations. This suggests that, at least in principle, there is no reason why active inference should only work with expected free energy as an intrinsic motivation. Finally, in Section 8 formulate active inference for arbitrary action-value functions which include those induced by intrinsic motivations. Modifying the generative model of Section 6.1 and looking at the variational approximation of its posterior comes close but does not correspond to the original active inference of Friston et al. (2015). We explain the additional trick that is needed.

In the Appendix we provide some more detailed calculations as well as notation translation tables (**Appendix C**) from our own to those of Friston et al. (2015) and Friston et al. (2016b).

4. NOTATION

We will explain our notation in more detail in the text, but for readers that mostly look at equations we give a short summary. Note that, **Appendix C** comprises a translation between Friston et al. (2015, 2016b) and the present notation. Mostly, we will denote random variables by upper case letters e.g., $X, Y, A, E, M, S, ...$ their state spaces by calligraphic upper case letters $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{E}, \mathcal{M}, \mathcal{S}...$, specific values of random variables which are elements of the state spaces by lower case letters $x, y, a, e, m, s, ...$ An exception to this are random variables that act as parameters of probability distributions. For those, we use upper case Greek letters $\Xi, \Phi, \Theta, ...$, for their usually continuous state spaces we use $\Delta_\Xi, \Delta_\Theta, \Delta_\Phi, ...$ and for specific values the lower case Greek letters $\xi, \phi, \theta, ...$ In cases where a random variable plays the role of an estimate of another variable X, we write the estimate as $\hat{X}$, its state space as $\hat{\mathcal{X}}$ and its values as $\hat{x}$.

We distinguish different types of probability distributions with letters p, q, r, and d. Here, p corresponds to probability distributions describing properties of the physical world including the agent and its environment, q identifies model probabilities used by the agent internally, r denotes approximations of such model probabilities which are also internal to the agent, and d denotes a probability distribution that can be replaced by a q or a r distribution. We write conditional probabilities in the usual way, e.g., $\mathrm{p}(y|x)$. For a model of this conditional probability parameterized by θ, we write $\mathrm{q}(\hat{y}|\hat{x}, \theta)$.

5. PERCEPTION-ACTION LOOP

In this section we introduce an agent's perception-action loop (PA-loop) as a causal Bayesian network. This formalism forms the basis for our treatment of active inference. The PA-loop should be seen as specifying the (true) dynamics of the underlying physical system that contains agent and environment as well as their interactions. In Friston's formulation, the environment dynamics of the PA-loop are referred to as the *generative process*. In general these dynamics are inaccessible to the agent itself. Nonetheless, parts of these (true) dynamics are often assumed to be known to the agent in order to simplify computation (see e.g., Friston et al., 2015). We first formally introduce the PA-loop as causal Bayesian network, and then state specific assumptions for the rest of this article.

5.1. PA-loop Bayesian Network

Figure 1 shows an agent's PA-loop, formalized as causal Bayesian network. The network describes the following causal dependencies over time: At $t = 0$ an initial environment state $e_0 \in \mathcal{E}$ leads to an initial sensor value $s_0 \in \mathcal{S}$. This sensor value influences the memory state $m_1 \in \mathcal{M}$ of the agent at time $t = 1$. Depending on this memory state, action $a_1 \in \mathcal{A}$ is performed which influences the transition of the environment state from e_0 to $e_1 \in \mathcal{E}$. The new environment state leads to a new sensor value s_1 which, together with the performed action a_1 and the memory state m_1, influence the next memory state m_2. The loop then continues in this way until a final time step T.

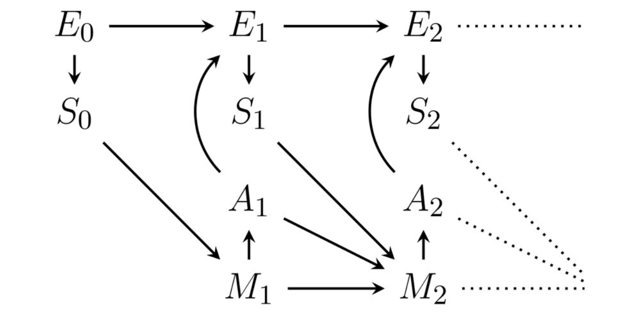


FIGURE 1 | First two time steps of the Bayesian network representing the perception-action loop (PA-loop). All subsequent time steps are identical to the one from time $t = 1$ to $t = 2$.

We assume that all variables are finite and that the PA-loop is time-homogeneous[1]. We exclude the first transition from $t = 0$ to $t = 1$ from the assumption of time-homogeneity in order to avoid having to pick an arbitrary action which precedes the investigated time-frame. The first transition is thus simplified to $p(m_1|s_0, a_0) := p(m_1|s_0)$. Under the assumption of time-homogeneity and the causal dependencies expressed in **Figure 1**, the joint probability distribution over the entire PA-loop is defined by:

$$p(e_{0:T}, s_{0:T}, a_{1:T}, m_{1:T}) = \left(\prod_{t=1}^{T} p(a_t|m_t)\, p(m_t|s_{t-1}, a_{t-1})\, p(s_t|e_t) \times p(e_t|a_t, e_{t-1})\right) p(s_0|e_0)\, p(e_0) \qquad (1)$$

where $e_{0:T}$ is shorthand for states $(e_0, e_1, \ldots, e_T)$. In order to completely determine this distribution we therefore have to specify the state spaces $\mathcal{E}, \mathcal{S}, \mathcal{A}$, and $\mathcal{M}$ as well as the following probabilities and mechanisms for all $e_0, e_t, e_{t+1} \in \mathcal{E}$; $s_0, s_t \in \mathcal{S}$; $a_t, a_{t+1} \in \mathcal{A}$; $m_1, m_t, m_{t+1} \in \mathcal{M}$ for $t > 0$:

- Initial environment distribution: $p(e_0)$,
- Environment dynamics: $p(e_{t+1}|a_{t+1}, e_t)$,
- Sensor dynamics: $p(s_t|e_t)$,
- Action generation: $p(a_t|m_t)$,
- Initial memory step $p(m_1|s_0)$,
- Memory dynamics: $p(m_{t+1}|s_t, a_t, m_t)$.

In the following we will refer to a combination of initial environment distribution, environment dynamics, and sensor dynamics simply as an *environment*. Similarly, an *agent* is a particular combination of initial memory step, memory dynamics, and action generation. The indexing convention we use here is identical to the one used for the generative model (see Section 6.1) in Friston et al. (2015).

Also, note the dependence of M_t on S_{t-1}, M_{t-1}, and additionally A_{t-1} in **Figure 1**. In the literature, the dependence on A_{t-1} is frequently not allowed (Ay et al., 2012; Ay and Löhr, 2015). However, we assume an efference-like update of the memory. Note that this dependence in addition to the dependence on m_{t-1} is only relevant if the actions are not deterministic functions of the memory state[2]. If action selection is probabilistic, knowing the outcome a_{t-1} of the action generation mechanism $p(a_{t-1}|m_{t-1})$ will convey more information than only knowing the past memory state m_{t-1}. This additional information can be used in inference about the environment state and fundamentally change the intrinsic perspective of an agent. We do not discuss these changes in more detail here but the reader should be aware of the assumption.

In a realistic robot scenario, the action a_t, if it is to be known by the agent, can only refer to the "action signal" or "action value" that is sent to the robot's physical actuators. These actuators will usually be noisy and the robot will not have access to the final effect of the signal it sends. The (noisy) conversion of an action signal to a physical configuration change of the actuator is here seen as part of the environment dynamics $p(e_t|a_t, e_{t-1})$. Similarly, the sensor value is the signal that the physical sensor of the robot produces as a result of a usually noisy measurement, so just like the actuator, the conversion of a physical sensor configuration to a sensor value is part of the sensor dynamics $p(s_t|e_t)$ which in turn belongs to the environment. As we will see later, the actions and sensor values must have well-defined state spaces $\mathcal{A}$ and $\mathcal{S}$ for inference on an internal model to work. This further justifies this perspective.

5.2. Assumptions

For the rest of this article we assume that the environment state space $\mathcal{E}$, sensor state space $\mathcal{S}$ as well as environment dynamics $p(e_{t+1}|a_{t+1}, e_t)$ and sensor dynamics $p(s_t|e_t)$ are arbitrarily fixed and that some initial environmental state e_0 is given. Since we are interested in intrinsic motivations, our focus is not on specific environment or sensor dynamics but almost exclusively on action generation mechanisms of agents that rely minimally on the specifics of these dynamics.

In order to focus on action generation, we assume that all the agents we deal with here have the same memory dynamics. For this, we choose a memory that stores all past sensor values $s_{\prec t} = (s_0, s_1, ..., s_{t-1})$ and actions $a_{\prec t} = (a_1, a_2, ..., a_{t-1})$ in the memory state m_t. This type of memory is also used in Friston et al. (2015, 2016b) and provides the agent with all existing data about its interactions with the environment. In this respect, it could be called a perfect memory. At the same time, whatever the agent learned from $s_{\prec t}$ and $a_{\prec t}$ that remains true based on the next time step's $s_{\preceq t+1}$ and $a_{\preceq t+1}$ must be relearned from scratch by the agent. A more efficient memory use might store only a sufficient statistic of the past data and keep reusable results of computations in memory. Such improvements are not part of this article (see e.g., Fox and Tishby, 2016, for discussion).

Formally, the state space $\mathcal{M}$ of the memory is the set of all sequences of sensor values and actions that can occur. Since there

[1] This means that all state spaces and transition probabilities are independent of the time step, e.g., $\mathcal{M}_t = \mathcal{M}_{t-1}$ and $p(s_t|e_t) = p(s_{t-1}|e_{t-1})$.

[2] In the deterministic case there is a function $f : \mathcal{M} \rightarrow \mathcal{A}$ such that $p(m_t|s_{t-1}, a_{t-1}, m_{t-1}) = p(m_t|s_{t-1}, f(m_{t-1}), m_{t-1}) = p(m_t|s_{t-1}, m_{t-1})$.

is only a sensor value and no action at $t = 0$, these sequences always begin with a sensor value followed by pairs of sensor values and actions. Furthermore, the sensor value and action at $t = T$ are never recorded. Since we have assumed a time-homogeneous memory state space $\mathcal{M}$ we must define it so that it contains all these possible sequences from the start. Formally, we therefore choose the union of the spaces of sequences of a fixed length (similar to a Kleene-closure):

$$\mathcal{M} = \mathcal{S} \cup \left(\bigcup_{t=1}^{T-1} \mathcal{S} \times (\mathcal{S} \times \mathcal{A})^t \right). \tag{2}$$

With this we can define the dynamics of the memory as:

$$\mathrm{p}(m_1|s_0) := \begin{cases} 1 & \text{if } m_1 = s_0 \\ 0 & \text{else.} \end{cases} \tag{3}$$

$$\mathrm{p}(m_t|s_{t-1}, a_{t-1}, m_{t-1}) := \begin{cases} 1 & \text{if } m_t = m_{t-1}s_{t-1}a_{t-1} \\ 0 & \text{else.} \end{cases} \tag{4}$$

This perfect memory may seem unrealistic and can cause problems if the sensor state space is large (e.g., high resolution images). However, we are not concerned with this type of problem here. Usually, the computation of actions based on past actions and sensor values becomes a challenge of efficiency long before storage limitations kick in: the necessary storage space for perfect memory only increases linearly with time, while, as we show later, the number of operations for Bayesian inference increases exponentially.

For completeness we also note how the memory dynamics look if actions are a deterministic function $f : \mathcal{M} \to \mathcal{A}$ of the memory state. Recall that in this case we can drop the edge from A_{t-1} to M_t in the PA-loop in **Figure 1** and have $a_t = f(m_t)$ so that we can define:

$$\mathrm{p}(m_1|s_0) := \begin{cases} 1 & \text{if } m_1 = s_0 \\ 0 & \text{else.} \end{cases} \tag{5}$$

$$\mathrm{p}(m_t|s_{t-1}, m_{t-1}) := \begin{cases} 1 & \text{if } m_t = m_{t-1}s_{t-1}f(m_{t-1}) \\ 0 & \text{else.} \end{cases} \tag{6}$$

Given a fixed environment and the memory dynamics, we only have to define the action generation mechanism $\mathrm{p}(a_t|m_t)$ to fully specify the perception-action loop. This is the subject of the next two sections.

In order to stay as close to Friston et al. (2015) as possible, we first explain the individual building blocks that can be extracted from Friston's active inference as described in Friston et al. (2015). These are the variational inference and the action selection. We then show how these two building blocks are combined in the original formulation. We eventually leverage our separation of components to show how the action selection component can be modified, and thus extend the active inference framework.

6. INFERENCE AND COMPLETE POSTERIORS

Ultimately, an agent needs to select actions. Inference based on past sensor values and actions is only needed if it is relevant to the action selection. Friston's active inference approach promises to perform action selection within the same inference step that is used to update the agent's model of the environment. In this section, we look at the inference component only and show how an agent can update a generative model in response to observed sensor values and performed actions.

The natural way of updating such a model is Bayesian inference via Bayes' rule. This type of inference leads to what we call the *complete posterior*. The complete posterior represents all knowledge that the agent can obtain about the consequences of its actions from its past sensor values and actions. In Section 7 we discuss how the agent can use the complete posterior to decide what is the best action to take.

Bayesian inference as straightforward recipe is usually not practical due to computational costs. The memory requirements of the complete posterior update increases exponentially with time and so does the number of operations needed to select actions. To keep the computational tractable, we have to limit ourselves to only use parts of the complete posterior. Furthermore, since the direct expressions (even of parts) of complete posteriors are usually intractable, approximations are needed. Friston's active inference is committed to variational inference as an approximation technique. Therefore, we explain how variational inference can be used as an approximation technique. Our setup for variational inference (generative model and approximate posterior) is identical to the one in Friston et al. (2015), but in this section we ignore the inference of actions included there. We will look at the extension to action inference in Section 7.

In the perception-action loop in **Figure 1**, action selection (and any inference mechanism used in the course of it) depends exclusively on the memory state m_t. As mentioned in Section 5, we assume that this memory state contains all *past* sensor values $s_{\prec t}$ and all *past* actions $a_{\prec t}$. To save space, we write $sa_{\prec t} := (s_{\prec t}, a_{\prec t})$ to refer to both sensor values and actions. We then have:

$$m_t = sa_{\prec t}. \tag{7}$$

However, since it is more intuitive to understand inference with respect to past sensor values and actions than in terms of memory, we use $sa_{\prec t}$ explicitly here in place of m_t.

6.1. Generative Model

The inference mechanism, internal to the action selection mechanism $\mathrm{p}(a|m)$, takes place on a hierarchical generative model (or density, in the continuous case). "Hierarchical" means that the model has parameters and hyperparameters, and "generative" indicates that the model relates *parameters and latent variables*, i.e., the environment state, as "generative" causes to sensor values and actions as *data* in a joint distribution. The generative model we investigate here is a part of the generative model used in Friston et al. (2015). For now, we omit the probability distribution over future actions and the "precision", which are only needed for active inference and are discussed later. The generative models in Friston et al. (2016a,b, 2017a) are all closely related.

Note that we are not inferring the causal structure of the Bayesian network or state space cardinalities, but define the generative model as a fixed Bayesian network with the graph shown in **Figure 2**. It is possible to infer the causal structure (see e.g., Ellis and Wong, 2008), but in that case, it becomes impossible to represent the whole generative model as a single Bayesian network (Ortega, 2011).

The variables in the Bayesian network in **Figure 2** that model variables occurring outside of $\mathrm{p}(a|m)$ in the perception-action loop (**Figure 1**), are denoted as hatted versions of their counterparts. More precisely:

- $\hat{s} \in \hat{\mathcal{S}} = \mathcal{S}$ are modelled sensor values,
- $\hat{a} \in \hat{\mathcal{A}} = \mathcal{A}$ are modelled actions,
- $\hat{e} \in \hat{\mathcal{E}}$ are modelled environment states.

To clearly distinguish the probabilities defined by the generative model from the true dynamics, we use the symbol q instead of p. In accordance with **Figure 2**, and also assuming time-homogeneity, the joint probability distribution over all variables in the model until some final modelled time $\hat{T}$ is given by:

$$\begin{aligned}
&\mathrm{q}(\hat{e}_{0:T}, \hat{s}_{0:T}, \hat{a}_{1:T}, \theta^1, \theta^2, \theta^3, \xi^1, \xi^2, \xi^3) \\
&\quad := \left(\prod_{t=1}^{T} \mathrm{q}(\hat{s}_t|\hat{e}_t, \theta^1)\, \mathrm{q}(\hat{e}_t|\hat{a}_t, \hat{e}_{t-1}, \theta^2)\, \mathrm{q}(\hat{a}_t)\right) \\
&\qquad \times \mathrm{q}(\hat{s}_0|\hat{e}_0, \theta^1)\, \mathrm{q}(\hat{e}_0|\theta^3) \left(\prod_{i=1}^{3} \mathrm{q}(\theta^i|\xi^i)\, \mathrm{q}(\xi^i)\right) \qquad (8)
\end{aligned}$$

Here, $\theta^1, \theta^2, \theta^3$ are the parameters of the hierarchical model, and ξ^1, ξ^2, ξ^3 are the hyperparameters. To save space, we combine the parameters and hyperparameters by writing

$$\theta := (\theta^1, \theta^2, \theta^3) \qquad (9)$$

$$\xi := (\xi^1, \xi^2, \xi^3). \qquad (10)$$

To fully specify the generative model, or equivalently a probability distribution over **Figure 2**, we have to specify the state spaces $\hat{\mathcal{E}}, \hat{\mathcal{S}}, \hat{\mathcal{A}}$ and:

- $\mathrm{q}(\hat{s}|\hat{e}, \theta^1)$ the sensor dynamics model,
- $\mathrm{q}(\hat{e}'|\hat{a}', \hat{e}, \theta^2)$ the environment dynamics model,
- $\mathrm{q}(\hat{e}_0|\theta^3)$ the initial environment state model,
- $\mathrm{q}(\theta^1|\xi^1)$ the sensor dynamics prior,
- $\mathrm{q}(\theta^2|\xi^2)$ the environment dynamics prior,
- $\mathrm{q}(\theta^3|\xi^3)$ the initial environment state prior,
- $\mathrm{q}(\xi^1)$ sensor dynamics hyperprior,
- $\mathrm{q}(\xi^2)$ environment dynamics hyperprior,
- $\mathrm{q}(\xi^3)$ initial environment state hyperprior,
- $\hat{T}$ last modelled time step,
- $\mathrm{q}(\hat{a}_t)$ for all $t \in \{1, , ..., \hat{T}\}$ the probability distribution over the actions at time t.

The state spaces of the parameters and hyperparameters are determined by the choice of $\hat{\mathcal{E}}, \hat{\mathcal{S}}, \hat{\mathcal{A}}$. We will see in Section 6.2 that $\hat{\mathcal{S}} = \mathcal{S}$ and $\hat{\mathcal{A}} = \mathcal{A}$ should be chosen in order to use this model for inference on past sensor values and actions. For $\hat{\mathcal{E}}$ it is not necessary to set it equal to $\mathcal{E}$ for the methods described to work. We note that if we set $\hat{\mathcal{E}}$ equal to the memory state space of Equation (2) the model and its updates become equivalent to those used by the Bayesian universal reinforcement learning agent Hutter (2005) in a finite (environment and time-interval) setting (see Section 6.3).

The last modelled time step $\hat{T}$ can be chosen as $\hat{T} = T$, but it is also possible to always set it to $\hat{T} = t + n$, in which case n specifies a future time horizon from current time step t. Such an agent would model a future that goes beyond the externally specified last time step T. The dependence of $\hat{T}$ on t (which we do not denote explicitly) within $\mathrm{p}(a|m)$ is possible since the current time step t is accessible from inspection of the memory state m_t which contains a sensor sequence of length t.

The generative model assumes that the actions are not influenced by any other variables, hence we have to specify action probabilities. This means that the agent does not model how its actions come about, i.e., it does not model its own decision process. Instead, the agent is interested in the (parameters of) the environment and sensor dynamics. It actively sets the probability distributions over past and future actions according to its needs. In practice, it either fixes the probability distributions to particular values (by using Dirac delta distributions) or to values that optimize some measure. We look into the optimization options in more detail later.

Note that the parameters and hyperparameters are standard random variables in the Bayesian network of the model. Also, the rules for calculating probabilities according to this model are just the rules for calculating probabilities in this Bayesian network.

In what follows, we assume that the hyperparameters are fixed as $\Xi^1 = \xi^1, \Xi^2 = \xi^2, \Xi^3 = \xi^3$. The following

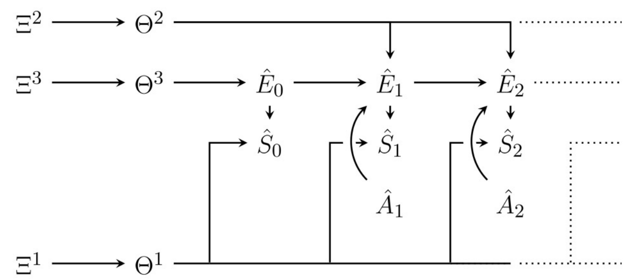


FIGURE 2 | Bayesian network of the generative model with parameters $\mathbf{\Theta} = (\mathbf{\Theta}^1, \mathbf{\Theta}^2, \mathbf{\Theta}^3)$ and hyperparameters $\Xi = (\Xi^1, \Xi^2, \Xi^3)$. Hatted variables are models / estimates of non-hatted counterparts in the perception-action loop in **Figure 1**. An edge that splits up connecting one node to n nodes (e.g., $\mathbf{\Theta}^2$ to $\hat{E}_1, \hat{E}_2, ...$) corresponds to n edges from that node to all the targets under the usual Bayesian network convention. Note that in contrast to the perception-action loop in **Figure 1**, imagined actions $\hat{A}_t$ have no parents. They are either set to past values or, for those in the future, a probability distribution over them must be assumed.

procedures (including both Bayesian and variational inference) can be generalized to also infer hyperparameters. However, our main reference (Friston et al., 2015) and most publications on active inference also fix the hyperparameters.

6.2. Bayesian Complete Posteriors

During action generation [i.e., within $\mathrm{p}(a|m)$] at time t, the agent has retained all its previously perceived sensor states and its previously performed actions in memory. The "experience" or data contained in its memory is thus $m_t = sa_{\prec t}$. This data can be plugged into the generative model to obtain posterior probability distributions over all non-observed random variables. Also, the model can estimate the not yet observed sensor values $\hat{s}_{t:\hat{T}}$, past and future unobservable environment states $\hat{e}_{0:\hat{T}}$, parameters θ and hyperparameters ξ. These estimations are done by setting:

$$\hat{A}_\tau = a_\tau, \text{for } \tau < t \tag{11}$$

and

$$\hat{S}_\tau = s_\tau, \text{for } \tau < t. \tag{12}$$

as shown in **Figure 3** for $t = 2$. For these assignments to be generally possible, we need to choose $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$ equal to $\mathcal{A}$ and $\mathcal{S}$ respectively. The resulting posterior probability distribution over all non-observed random variables is then, according to standard rules of calculating probabilities in a Bayesian network:

$$\begin{aligned}&\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi)\\ &\quad := \frac{\mathrm{q}(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)}{\int \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}} \mathrm{q}(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)\, \mathrm{d}\theta}.\end{aligned} \tag{13}$$

Eventually, the agent needs to evaluate the consequences of its future actions. Just as it can update the model with respect to past actions and sensor values, the agent can update its evaluations with "contemplated" future action sequences $\hat{a}_{t:\hat{T}}$. For each such future action sequence $\hat{a}_{t:\hat{T}}$, the agent obtains a distribution over the remaining random variables in the model:

$$\begin{aligned}&\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)\\ &\quad := \frac{\mathrm{q}(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)}{\int \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}} \mathrm{q}(s_{\prec t}, \hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, a_{\prec t}, \hat{a}_{t:\hat{T}}, \theta, \xi)\, \mathrm{d}\theta}.\end{aligned} \tag{14}$$

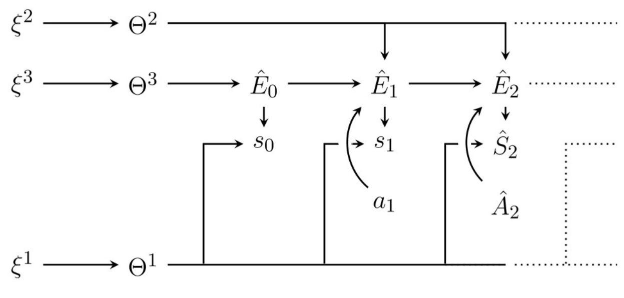


FIGURE 3 | Internal generative model with plugged in data up to $t = 2$ with $\hat{S}_0 = s_0, \hat{S}_1 = s_1$ and $\hat{A}_1 = a_1$ as well as from now on fixed hyperparameters $\xi = (\xi^1, \xi^2, \xi^3)$. Conditioning on the plugged in data leads to the posterior distribution $\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi)$. Predictions for future sensor values can be obtained by marginalising out other random variables e.g., to predict $\hat{S}_2$ we would like to get $\mathrm{q}(\hat{s}_2 | s_0, s_1, a_1, \xi)$. Note however that this requires an assumption for the probability distribution over $\hat{A}_2$.

We call each such distribution a *Bayesian complete posterior*. We choose the term complete posterior since the "posterior" by itself usually refers to the posterior distribution over the parameters and latent variables $\mathrm{q}(\theta, \hat{e}_{t-1} | sa_{\prec t}, \xi)$ [we here call this a *posterior factor*, see Equation (16)] and the posterior predictive distributions marginalize out the parameters and latent variables to get $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The complete posteriors are probability distributions over all random variables in the generative model including parameters, latent variables, and future variables. In this sense the set of all (Bayesian) complete posteriors represents the complete knowledge state of the agent at time t about consequences of future actions after updating the model with past actions and observed sensor values $sa_{\prec t}$. At each time step the sequence of past actions and sensor values is extended from $sa_{\prec t}$ to $sa_{\prec t+1}$ (i.e., m_t goes to m_{t+1}) and a new set of complete posteriors is obtained.

All intrinsic motivations discussed in this article evaluate future actions based on quantities that can be derived from the corresponding complete posterior.

It is important to note that the complete posterior can be factorized into a term containing the influence of past sensor values and actions (data). This factorization can be made on the parameters θ and ξ, the environment states $\hat{e}_{\prec t}$, predicted future environment states $\hat{e}_{t:\hat{T}}$ and sensor values $\hat{s}_{t:\hat{T}}$ depending on the future actions $\hat{a}_{t:\hat{T}}$, and the estimated environment state $\hat{e}_{t-1}$ and θ. Using the conditional independence

$$SA_{\prec t} \perp\!\!\!\perp \hat{S}_{t:\hat{T}}, \hat{E}_{t:\hat{T}} \mid \hat{A}_{t:\hat{T}}, \hat{E}_{t-1}, \Theta, \Xi, \tag{15}$$

which can be identified (via *d*-separation; Pearl, 2000) from the Bayesian network in **Figure 3**, we can rewrite this as:

$$\begin{aligned}&\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)\\ &\quad = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi).\end{aligned} \tag{16}$$

This equation represents the desired factorization. This formulation separates complete posteriors into a predictive and a posterior factor. The predictive factor is given as part of the generative model (Equation 8)

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) = \prod_{r=t}^{\hat{T}} \mathrm{q}(\hat{s}_r | \hat{e}_r, \theta^1)\, \mathrm{q}(\hat{e}_r | \hat{a}_r, \hat{e}_{r-1}, \theta^2) \tag{17}$$

and does not need to be updated through calculations at different time steps. This factor contains the dependence of the complete posterior on future actions. This dependency reflects that, under the given generative model, the consequences of actions for each combination of Θ and $\hat{E}_{t-1}$ remain the same irrespective of experience. What changes when a new action and sensor value pair comes in is the distribution over the values of Θ and $\hat{E}_{t-1}$ and with them the *expectations* over consequences of actions.

On the other hand, the posterior factor must be updated at every time step. In **Appendix A**, we sketch the computation which shows that it involves a sum over $|\mathcal{E}|^t$ elements. This calculation is intractable as time goes on and one of the reasons to use approximate inference methods like variational inference.

Due to the above factorization, we may only need to approximate the posterior factor $\mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi)$ and use the exact predictive factor if probabilities involving future sensor values or environment states are needed.

This is the approach taken e.g., in Friston et al. (2015). However, it is also possible to directly approximate parts of the complete posterior involving random variables in both factors, e.g., by approximating $\mathrm{q}(\hat{e}_{0:\hat{T}}, \theta^1 | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. This latter approach is taken in Friston et al. (2016b) and we see it again in Equation (43) but in this publication the focus is on the former approach.

In the next section, we look at the special case of universal reinforcement learning before we go on to variational inference to approximate the posterior factor of the (Bayesian) complete posteriors.

6.3. Connection to Universal Reinforcement Learning

In this section, we relate the generative model of Equation (8) and its posterior predictive distribution to those used by the Bayesian universal reinforcement learning agent. Originally, this agent is defined by Hutter (2005). More recent work includes Leike (2016) and (for the current purpose sufficient and particularly relevant) Aslanides et al. (2017).

Let us set $\hat{\mathcal{E}} = \mathcal{M}$ with $\mathcal{M}$ as in Equation (2) and let the agent identify each past $sa_{\prec t}$ with a state of the environment, i.e.,

$$\hat{e}_{t-1} = sa_{\prec t}. \tag{18}$$

Under this definition the next environment state $\hat{e}_t$ is just the concatenation of the last environment state $sa_{\prec t}$ with the next next action selected by the agent $\hat{a}_t$ and the next sensor value $\hat{s}_t$:

$$\hat{e}_t = \hat{s}\hat{a}_{\preceq t} = sa_{\prec t}\hat{s}\hat{a}_t. \tag{19}$$

So given a next contemplated action $\bar{\hat{a}}_t$ the next environment state $\hat{e}_t$ is already partially determined. What remains to be predicted is only the next sensor value $\hat{s}_t$. Formally, this is reflected in the following derivation:

$$\mathrm{q}(\hat{e}_t | \bar{\hat{a}}_t, \hat{e}_{t-1}, \theta^2) := \mathrm{q}(\hat{s}_t, \hat{a}_t, \hat{s}\hat{a}_{\prec t} | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2) \tag{20}$$

$$= \mathrm{q}(\hat{s}_t | \hat{a}_t, \hat{s}\hat{a}_{\prec t}, \bar{\hat{a}}_t, sa_{\prec t}, \theta^2)\, \mathrm{q}(\hat{a}_t, \hat{s}\hat{a}_{\prec t} | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2) \tag{21}$$

$$= \mathrm{q}(\hat{s}_t | \hat{a}_t, \hat{s}\hat{a}_{\prec t}, \bar{\hat{a}}_t, sa_{\prec t}, \theta^2)\delta_{\bar{\hat{a}}_t}(\hat{a}_t)\delta_{sa_{\prec t}}(\hat{s}\hat{a}_{\prec t}) \tag{22}$$

$$= \mathrm{q}(\hat{s}_t | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2)\delta_{\bar{\hat{a}}_t}(\hat{a}_t)\delta_{sa_{\prec t}}(\hat{s}\hat{a}_{\prec t}). \tag{23}$$

This shows that in this case the model of the next environment state (the left hand side) is determined by the model of the next sensor value $\mathrm{q}(\hat{s}_t | \bar{\hat{a}}_t, sa_{\prec t}, \theta^2)$.

So instead of carrying a distribution over possible models of the next environment state such an agent only needs to carry a distribution over models of the next sensor value. Furthermore, an additional model $\mathrm{q}(\hat{s} | \hat{e}, \theta^1)$ of the dependence of the sensor values on environment states parameterized by θ^1 is superfluous. The next predicted sensor value is already predicted by the model $\mathrm{q}(\hat{s}_t | \hat{a}_t, sa_{\prec t}, \theta^2)$. It is therefore possible to drop the parameter θ^1.

The parameter θ^3, for the initial environment state distribution, becomes a distribution over the initial sensor value since $\hat{e}_0 = \hat{s}_0$:

$$\mathrm{q}(\hat{e}_0 | \theta^3) = \mathrm{q}(\hat{s}_0 | \theta^3). \tag{24}$$

We can then derive the posterior predictive distribution and show that it coincides with the one given in Aslanides et al. (2017). For the complete posterior of Equation (16) we find:

$$\begin{aligned}
&\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \\
&\quad = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) && \text{(16 revisited)} \\
&\quad = \mathrm{q}(\hat{e}_{t:\hat{T}} | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) && (25) \\
&\quad = \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \theta)\, \mathrm{q}(\theta | sa_{\prec t}, \xi) \\
&\qquad \times \prod_{\tau=0}^{t} \delta_{sa_{\prec \tau}}(\hat{e}_\tau) \prod_{\tau=t+1}^{\hat{T}} \delta_{sa_{\prec t}\hat{s}\hat{a}_{t:\tau}}(\hat{e}_\tau). && (26)
\end{aligned}$$

To translate this formulation into the notation of Aslanides et al. (2017) first drop the representation of the environment state which is determined by the sensor values and actions anyway. This means that the complete posterior only needs to predict future sensor values and parameters. Formally, this means the complete posterior can be replaced without loss of generality:

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \rightarrow \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \theta)\, \mathrm{q}(\theta | sa_{\prec t}, \xi). \tag{27}$$

To translate notations let $\theta \rightarrow \nu$; $\hat{a}, a \rightarrow a$; $\hat{s}, s \rightarrow e$. Also, set $\hat{T} \rightarrow t$ because only one step futures are considered in universal reinforcement learning (this is due to the use of policies instead of future action sequences). Then, the equation for the posterior predictive distribution

$$\mathrm{q}(\hat{s}_t | \hat{a}_t, sa_{\prec t}, \xi) = \int \mathrm{q}(\hat{s}_t | \hat{a}_t, sa_{\prec t}, \theta)\, \mathrm{q}(\theta | sa_{\prec t}, \xi)\, \mathrm{d}\theta, \tag{28}$$

is equivalent to Aslanides et al. (2017, Equation 5) (the sum replaces the integral for a countable Δ_Θ):

$$\xi(e | ae_{\prec t}, a) = \sum_\nu p(e | \nu, ae_{\prec t}, a) p(\nu | ae_{\prec t}) \tag{29}$$

$$\Leftrightarrow \xi(e) = \sum_\nu p(e | \nu) p(\nu), \tag{30}$$

where we dropped the conditioning on $ae_{\prec t}, a$ from the notation in the second line as done in the original (where this is claimed to improve clarity). Also note that $\xi(e)$ would be written $\mathrm{q}(e | \xi)$ in our notation. In the universal reinforcement learning literature parameters like θ (or ν) and ξ are sometimes directly used to denote the probability distribution that they parameterize.

Updating of the posterior $q(\theta|sa_{\prec t},\xi)$ in response to new data also coincides with updating of the weights $p(\nu)$:

$$
\begin{aligned}
q(\theta|sa_{\preceq t},\xi) &= \frac{q(\theta,s_t|a_t,sa_{\prec t},\xi)}{q(s_t|a_t,sa_{\prec t},\xi)} && (31)\\
&= \frac{q(s_t|a_t,sa_{\prec t},\theta,\xi)\,q(\theta|a_t,sa_{\prec t},\xi)}{q(s_t|a_t,sa_{\prec t},\xi)} && (32)\\
&= \frac{q(s_t|a_t,sa_{\prec t},\theta)\,q(\theta|sa_{\prec t},\xi)}{q(s_t|a_t,sa_{\prec t},\xi)} && (33)\\
&= \frac{q(s_t|a_t,sa_{\prec t},\theta)}{q(s_t|a_t,sa_{\prec t},\xi)}\,q(\theta|sa_{\prec t},\xi). && (34)
\end{aligned}
$$

The first two lines are general. From the second to third we used

$$S_t \perp\!\!\!\perp \Xi|A_t,SA_{\prec t},\Theta \tag{35}$$

and

$$\Theta \perp\!\!\!\perp A_t|SA_{\prec t},\Xi \tag{36}$$

which follow from the Bayesian network structure **Figure 2**. In the notation of Aslanides et al. (2017) Equation (34) becomes

$$p(\nu|e) = \frac{p(e|\nu)}{p(e)}p(\nu). \tag{37}$$

This shows that assuming the same model class Δ_Θ the predictions and belief updates of an agent using the Bayesian complete posterior of Section 6.2 are the same as those of the Bayesian universal reinforcement learning agent. Action selection can then be performed just as in Aslanides et al. (2017) as well. This is done by selecting policies. In the present publication we instead select action sequences directly. However, in both cases the choice maximizes the value predicted by the model. More on this in Section 7.2.

6.4. Approximate Complete Posteriors

As mentioned in the last section, the complete posterior can be approximated via variational inference (see Attias, 1999; Winn and Bishop, 2005; Bishop, 2011; Blei et al., 2017). There are alternative methods such as belief propagation, expectation propagation (Minka, 2001; Vehtari et al., 2014), and sampling-based methods (Lunn et al., 2000; Bishop, 2011), but active inference commits to variational inference by framing inference as variational free energy minimization (Friston et al., 2015). Variational free energy (Equation 45) is just the negative evidence lower bound (ELBO) of standard variational inference (e.g., Blei et al., 2017). In the following, we show how the complete posterior can be approximated via variational inference.

The idea behind variational inference is to use a simple family of probability distributions and identify the member of that family which approximates the true complete posterior best. This turns inference into an optimization problem. According to Wainwright and Jordan (2007) this reformulation as an optimization problem is the essence of variational methods. If the family of distributions is chosen such that it includes the complete posterior then the optimization will eventually lead to the same result as Bayesian inference. However, one advantage of the formulation as an optimization is that it can also be performed over a family of probability distributions that is simpler than the family that includes the actual complete posterior. This is what turns variational inference into an approximate inference procedure. Usually, the (simpler) families of probability distributions are chosen as products of independent distributions.

Recalling Equation (16), the complete posterior as a product of a predictive and a posterior factor is:

$$
\begin{aligned}
&q(\hat{s}_{t:\hat{T}},\hat{e}_{0:\hat{T}},\theta|\hat{a}_{t:\hat{T}},sa_{\prec t},\xi)\\
&\quad = q(\hat{s}_{t:\hat{T}},\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},\hat{e}_{t-1},\theta)\,q(\hat{e}_{\prec t},\theta|sa_{\prec t},\xi). \qquad \text{(16 revisited)}
\end{aligned}
$$

This product is the main object of interest. We want to approximate the formula with a probability distribution that lets us (tractably) calculate the posteriors required by a given intrinsic motivation, which can consequently be used for action selection.

As mentioned before, to approximate the complete posterior we here approximate only the posterior factor and use the given generative model's predictive factor as is done in Friston et al. (2015)[3] The approximate posterior factor is then combined with the exact predictive factor to get the approximate complete posterior. Let us write $r(\hat{e}_{\prec t},\theta|\phi)$ for the approximate posterior factor (**Figure 4**), defined as:

$$
\begin{aligned}
r(\hat{e}_{\prec t},\theta|\phi) &:= r(\hat{e}_{\prec t}|\phi^{E_{\prec t}})\,r(\theta|\phi) && (38)\\
&:= \prod_{\tau=0}^{t-1} r(\hat{e}_\tau|\phi^{E_\tau}) \prod_{i=1}^{3} r(\theta^i|\phi^i). && (39)
\end{aligned}
$$

As we can see it models each of the random variables that the posterior factor ranges over as independent of all others. This is called a *mean field* approximation. Then, the approximate complete posterior (**Figure 5**) is:

$$r(\hat{s}_{t:\hat{T}},\hat{e}_{0:\hat{T}},\theta|\hat{a}_{t:\hat{T}},\phi) := q(\hat{s}_{t:\hat{T}},\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},\hat{e}_{t-1},\theta)\,r(\hat{e}_{\prec t},\theta|\phi). \tag{40}$$

Note that the variational parameter absorbs the hyperparameter ξ as well as the past sensor values and actions $sa_{\prec t}$. The parameter does not absorb future actions which are part of the predictive factor. The dependence on future actions needs to be kept if we want to select actions using the approximate complete posterior.

[3] A close inspection of Friston et al. (2015, Equation 9) shows that the approximate complete posterior that ends up being evaluated by the action-value function is the one we discuss in Equation (40). It uses the predictive factor to get the probabilities $r(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}},\hat{e}_{t-1},\phi)$ of future environment states. However, the approximate posterior in Friston et al. (2015, Equation 10) uses a factorization of all future environment states like the one we give in Equation (43). The probabilities of future environment states in that posterior are not used anywhere in Friston et al. (2015). In principle, they could be used as is done in Friston et al. (2016b, Equation 2.6) where the complete posterior of Equation (43) is used in the action-value function. Both approaches are possible.

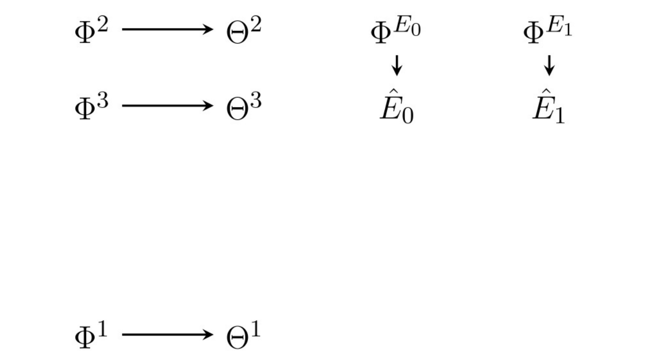


FIGURE 4 | Bayesian network of the approximate posterior factor at $t = 2$. The variational parameters Φ^1, Φ^2, Φ^3, and $\Phi^{E_{\prec t}} = (\Phi^{E_0}, \Phi^{E_1})$ are positioned so as to indicate what dependencies and nodes they replace in the generative model in **Figure 2**.

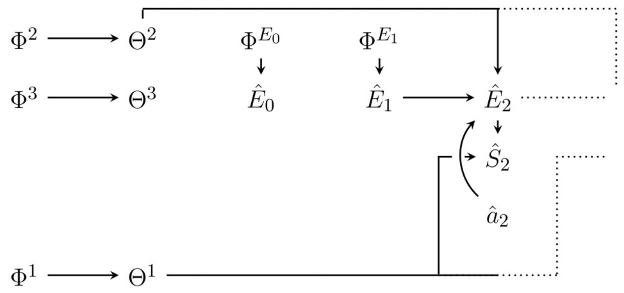


FIGURE 5 | Bayesian network of the approximate complete posterior of Equation (40) at $t = 2$ for the future actions $\hat{a}_{t:\hat{T}}$. Only $\hat{E}_{t-1}$, Θ^1, Θ^2 and the future action $\hat{a}_{t:\hat{T}}$ appear in the predictive factor and influence future variables. In general there is one approximate complete posterior for each possible sequence $\hat{a}_{t:\hat{T}}$ of future actions.

We have:

$$r(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, \phi) \approx q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \tag{41}$$

if

$$r(\hat{e}_{\prec t}, \theta|\phi) \approx q(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi). \tag{42}$$

This approximation can be achieved by standard variational inference methods.

For those interested more in the approximation of the complete posterior as in Friston et al. (2016b), we provide the used family of factorized distributions. It must be noted that the agent in this case carries a separate approximate posterior for each possible complete action sequence $\hat{a}_{0:T}$. For predictions of environment states, it does not use the predictive factor, but instead looks at the set of generative models compatible with the past. For each of those, the agent considers all environment states at different times as independent. The approximate posteriors, compatible with a past sequence of actions $a_{\prec t}$, are of the form:

$$\begin{aligned} &r(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta^1|\hat{a}_{t:\hat{T}}, a_{\prec t}, \phi^1) \\ &\quad = q(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, \theta^1) \prod_{\tau=0}^{\hat{T}} r(\hat{e}_\tau|\hat{a}_{t:\hat{T}}, a_{\prec t}, \phi^{E_\tau})\, r(\theta^1|\phi^1). \end{aligned} \tag{43}$$

Note also that the relation between sensor values and environment states is still provided by the generative models' sensor dynamics $q(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, \theta^1)$. In this article however, we focus on the approach in Friston et al. (2015) which requires only one approximate posterior at time t since future actions only occur in the predictive factors which we do not approximate.

We define the relative entropy (or KL-divergence) between the approximate and the true posterior factor:

$$\begin{aligned} &\mathrm{KL}[r(\hat{E}_{\prec t}, \Theta|\phi)||\, q(\hat{E}_{\prec t}, \Theta|sa_{\prec t}, \xi)] \\ &\quad := \sum_{\hat{e}_{\prec t}} \int r(\hat{e}_{\prec t}, \theta|\phi) \log \frac{r(\hat{e}_{\prec t}, \theta|\phi)}{q(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)}\, d\theta. \end{aligned} \tag{44}$$

Note that, we indicate the variables that are summed over by capitalizing them. The KL-divergence quantifies the difference between the two distributions. It is non-negative, and only zero if the approximate and the true posterior factor are equal (see e.g., Cover and Thomas, 2006).

The variational free energy, also known as the (negative) evidence lower bound (ELBO) in variational inference literature, is defined as:

$$\mathcal{F}[\xi, \phi, sa_{\prec t}] := \sum_{\hat{e}_{\prec t}} \int r(\hat{e}_{\prec t}, \theta|\phi) \log \frac{r(\hat{e}_{\prec t}, \theta|\phi)}{q(s_{\preceq t}, \hat{e}_{\prec t}, \theta|a_{\prec t}, \xi)}\, d\theta \tag{45}$$

$$\begin{aligned} &= -\log q(s_{\prec t}|a_{\prec t}, \xi) \\ &\quad + \mathrm{KL}[r(\hat{E}_{\prec t}, \Theta|\phi)||\, q(\hat{E}_{\prec t}, \Theta|sa_{\prec t}, \xi)] \end{aligned} \tag{46}$$

The first term in Equation (46) is the surprise of negative log evidence. For a fixed hyperparameter ξ it is a constant. Minimizing the variational free energy therefore directly minimizes the KL-divergence between the true and the approximate posterior factor given $sa_{\prec t}$ and ξ.

In our case, variational inference amounts to solve the optimization problem:

$$\phi^*_{sa_{\prec t},\xi} := \arg\min_{\phi} \mathcal{F}[\phi, sa_{\prec t}, \xi]. \tag{47}$$

This optimization is a standard problem. See Bishop (2011) and Blei et al. (2017) for ways to solve it.

The resulting variational parameters $\phi^*_{sa_{\prec t},\xi} = (\phi^{E_0}_{sa_{\prec t},\xi}, ..., \phi^{E_{t-1}}_{sa_{\prec t},\xi}, \phi^1{}_{sa_{\prec t},\xi}, \phi^2{}_{sa_{\prec t},\xi}, \phi^3{}_{sa_{\prec t},\xi})$ define the approximate posterior factor. The variational parameters, together with the exact predictive factors, allow us to compute the approximate complete posteriors for each sequence of future actions $\hat{a}_{t:\hat{T}}$:

$$\begin{aligned} &r(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, \phi^*_{sa_{\prec t},\xi}) \\ &\quad = q(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, r(\hat{e}_{\prec t}, \theta|\phi^*_{sa_{\prec t},\xi}) \end{aligned} \tag{48}$$

$$\approx q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi). \tag{49}$$

In the next section, we look at action selection as the second component of action generation. To this end, we show how to evaluate sequences of future actions $\hat{a}_{t:\hat{T}}$ by evaluating either Bayesian complete posteriors or the approximate complete posteriors.

7. ACTION SELECTION BASED ON INTRINSIC MOTIVATIONS

7.1. Intrinsic Motivation and Action-Value Functions

The previous section resulted in sets of Bayesian or approximate complete posteriors. Independently of whether a complete posterior is the approximate or the Bayesian version, it represents the entire knowledge of the agent about the consequences of the sequence of future actions $\hat{a}_{t:\hat{T}}$ that is associated with it. In order to evaluate sequences of future actions the agent can only rely on its knowledge which suggests that all such evaluations should depend solely on complete posteriors. One could argue that the motivation might also depend directly on the memory state containing $sa_{\prec t}$. We here take a position somewhat similar to the one proposed by Schmidhuber (2010) that intrinsic motivations concerns the "learning of a better world model." We consider the complete posterior as the current world model and assume that intrinsic motivations depend only on this model and not on the exact values of past sensor values and actions. As we will see this assumption is also enough to capture the three intrinsic motivations that we discuss here. This level of generality is sufficient for our purpose of extending the free energy principle. Whether it sufficient for a final and general intrinsic motivation definition is beyond the scope of this publication.

Complete posteriors are essentially conditional probability distributions over $\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta$ given elements of $\hat{\mathcal{A}}^{\hat{T}-t+1}$. A necessary (but not sufficient) requirement for intrinsic motivations in our context (agents with generative models) is then that they are functions on the space of such conditional probability distributions. Let $\Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1}\times\hat{\mathcal{E}}^{\hat{T}+1}\times\Delta_\Theta|\hat{\mathcal{A}}^{\hat{T}-t+1}}$ be the space of conditional probability distributions over $\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta$ given elements of $\hat{\mathcal{A}}^{\hat{T}-t+1}$. Then an *intrinsic motivation* is a function $\mathfrak{M} : \Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1}\times\hat{\mathcal{E}}^{\hat{T}+1}\times\Delta_\Theta|\hat{\mathcal{A}}^{\hat{T}-t+1}} \times \hat{\mathcal{A}}^{\hat{T}-t+1} \to \mathbb{R}$ taking a probability distribution $d(.,.,.|.) \in \Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1}\times\hat{\mathcal{E}}^{\hat{T}+1}\times\Delta_\Theta|\hat{\mathcal{A}}^{\hat{T}-t+1}}$ and a given future actions sequence $\hat{a}_{t:\hat{T}} \in \hat{\mathcal{A}}^{\hat{T}-t+1}$ to a real value $\mathfrak{M}(d(.,.,.|.), \hat{a}_{t:\hat{T}}) \in \mathbb{R}$. We can then see that the Bayesian complete posterior $q(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ for a fixed past $sa_{\prec t}$ written as $q(.,.,.|., sa_{\prec t}, \xi)$ provides such conditional probability distribution. Similarly, every member of the family of distributions used to approximate the Bayesian complete posterior via variational inference $r(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, \phi)$ written as $r(.,.,.|., \phi)$ also provides such a conditional probability distribution. It will become important when discussing active inference that the optimized value $\phi^*_{sa_{\prec t},\xi}$ of the variational parameters as well as any other value of the variational parameters ϕ define an element with the right structure to be evaluated together with a set of future actions by an intrinsic motivation function.

Using intrinsic motivation functions we then define two kinds of induced action-value functions. These are similar to value functions in reinforcement learning[4] The first is the *Bayesian action-value function* (or functional):

$$\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) := \mathfrak{M}(q(.,.,.|., sa_{\prec t}, \xi), \hat{a}_{t:\hat{T}}). \tag{50}$$

In words the Bayesian action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ infers the set of Bayesian complete posteriors of past experience $sa_{\prec t}$ and then evaluates the sequence of future actions $\hat{a}_{t:\hat{T}}$ according to the intrinsic motivation function $\mathfrak{M}$.

The *variational action-value function* is defined as[5]:

$$\hat{Q}(\hat{a}_{t:\hat{T}}, \phi) := \mathfrak{M}(r(.,.,.|., \phi), \hat{a}_{t:\hat{T}}). \tag{51}$$

So the variational action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$ directly takes the conditional probability distribution defined by variational parameter ϕ and evaluates the sequence of future actions $\hat{a}_{t:\hat{T}}$ according to $\mathfrak{M}$. Unlike in the Bayesian case no inference takes place during the evaluation of $\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$.

At the same time, after variational inference, if we plug in $\phi^*_{sa_{\prec t},\xi}$ for ϕ we have:

$$\hat{Q}(\hat{a}_{t:\hat{T}_a}, \phi^*_{sa_{\prec t},\xi}) \approx \hat{Q}(\hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi). \tag{52}$$

Note that the reason we have placed a hat on $\hat{Q}$ is that, even in the Bayesian case, it is usually not the optimal action-value function but instead is an estimate based on the current knowledge state represented by the complete posteriors of the agent.

Also note that some intrinsic motivations (e.g., empowerment) evaluate e.g., the next n actions by using predictions reaching $n + m$ steps into the future. This means that they need all complete posteriors for $\hat{a}_{t:t+n+m-1}$ but only evaluate the actions $\hat{a}_{t:t+n-1}$. In other words they cannot evaluate actions up to their generative model's time-horizon $\hat{T}$ but only until a shorter time-horizon $\hat{T}_a = \hat{T} - m$ for some natural number m. When necessary we indicate such a situation by only passing shorter future action sequences $\hat{a}_{t:\hat{T}_a}$ to the action-value function, in turn, the intrinsic motivation function. The respective posteriors keep the original time horizon $\hat{T} > \hat{T}_a$.

7.2. Deterministic and Stochastic Action Selection

We can then select actions simply by picking the first action in the sequence $\hat{a}_{t:\hat{T}}$ that maximizes the Bayesian action-value function:

$$\hat{a}^*_{t:\hat{T}}(m_t) := \hat{a}^*_{t:\hat{T}}(sa_{\prec t}) := \arg\max_{\hat{a}_{t:\hat{T}}} \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \tag{53}$$

[4] The main difference is that the action-value functions here evaluate sequences of future actions as opposed to policies. This is the prevalent practice in active inference literature including Friston et al. (2015) and we therefore follow it here.

[5] We abuse notation here by reusing the same symbol $\hat{Q}$ for the variational action-value function as for the Bayesian action-value function. However, in this publication the argument ($sa_{\prec t}, \xi$ or ϕ) always indicates which one is meant.

and set

$$\hat{a}^*(m_t) := \hat{a}^*_t(m_t). \tag{54}$$

or for the variational action value function:

$$\hat{a}^*_{t:\hat{T}}(m_t) := \hat{a}^*_{t:\hat{T}}(\phi^*_{sa_{\prec t},\xi}) := \arg\max_{\hat{a}_{t:\hat{T}}} \hat{Q}(\hat{a}_{t:\hat{T}}, \phi^*_{sa_{\prec t},\xi}). \tag{55}$$

and set

$$\hat{a}^*(m_t) := \hat{a}^*_t(m_t). \tag{56}$$

This then results in a deterministic action generation $\mathrm{p}(a|m)$:

$$\mathrm{p}(a_t|m_t) := \delta_{\hat{a}^*(m_t)}(a_t).$$

We note here that in the case of universal reinforcement learning the role of $\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ is played by $V^\pi_\xi(sa_{\prec t})$. There π is a policy that selects actions in dependence on the entire past $sa_{\prec t}$ and ξ parameterizes the posterior just like in the present publication. The arg max in Equation (53) selects a policy instead of an action sequence and that policy is used for the action generation.

A possible stochastic action selection that is important for active inference is choosing the action according to a so called softmax policy (Sutton and Barto, 1998):

$$\mathrm{p}(a_t|m_t) := \sum_{\hat{a}_{t+1:\hat{T}}} \frac{1}{Z(\gamma, sa_{\prec t}, \xi)} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)} \tag{57}$$

where:

$$Z(\gamma, sa_{\prec t}, \xi) := \sum_{\hat{a}_{t:\hat{T}}} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)} \tag{58}$$

is a normalization factor. Note that we are marginalizing out later actions in the sequence $\hat{a}_{t:\hat{T}}$ to get a distribution only over the action $\hat{a}_t$. For the variational action-value function this becomes:

$$\mathrm{p}(a_t|m_t) := \sum_{\hat{a}_{t+1:\hat{T}}} \frac{1}{Z(\gamma, \phi^*_{sa_{\prec t},\xi})} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, \phi^*_{sa_{\prec t},\xi})} \tag{59}$$

where:

$$Z(\gamma, \phi^*_{sa_{\prec t},\xi}) := \sum_{\hat{a}_{t:\hat{T}}} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, \phi^*_{sa_{\prec t},\xi})}. \tag{60}$$

Since it is relevant for active inference (see Section 8), note that the softmax distribution over future actions can also be defined for arbitrary ϕ and not only for the optimized $\phi^*_{sa_{\prec t},\xi}$. At the same time, the softmax distribution for the optimized $\phi_{sa_{\prec t},\xi}$ clearly also approximates the softmax distribution of the Bayesian action-value function.

Softmax policies assign action sequences with higher values of $\hat{Q}$ higher probabilities. They are often used as a replacement for the deterministic action selection to introduce some exploration. Here, lower γ leads to higher exploration; conversely, in the limit where $\gamma \to \infty$ the softmax turns into the deterministic action selection. From an intrinsic motivation point of view such additional exploration should be superfluous in many cases since many intrinsic motivations try to directly drive exploration by themselves. Another interpretation of such a choice is to see γ as a trade-off factor between the processing cost of choosing an action precisely and achieving a high action-value. The lower γ, the higher the cost of precision. This leads to the agent more often taking actions that do not attain maximum action-value.

We note that the softmax policy is not the only possible stochastic action selection mechanism. Another option discussed in the literature is Thompson sampling (Ortega and Braun, 2010, 2014; Aslanides et al., 2017). In our framework this corresponds to a two step action selection procedure where we first sample an environment and parameter pair $(\bar{\hat{e}}_{t-1}, \bar{\theta})$ from a posterior factor (Bayesian or variational)

$$(\bar{\hat{e}}_{t-1}, \bar{\theta}) \sim \mathrm{d}(\hat{E}_{t-1}, \Theta | sa_{\prec t}, \xi) \tag{61}$$

then plug the according predictive factor $\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \bar{\hat{e}}_{t-1}, \bar{\theta})$ into the action value function

$$\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) := \mathfrak{M}(\mathrm{q}(.,.|.,\bar{\hat{e}}_{t-1}, \bar{\theta}), \hat{a}_{t:\hat{T}}). \tag{62}$$

This allows intrinsic motivations that only evaluate the probability distribution over future sensor values $\hat{S}_{t:\hat{T}}$ and environment states $\hat{E}_{t:\hat{T}}$. However, it rules out those that evaluate the posterior probability of environment parameters Θ because we sample a specific $\bar{\theta}$.

7.3. Intrinsic Motivations

Now, we look at some intrinsic motivations including the intrinsic motivation part underlying Friston's active inference.

In the definitions, we use $\mathrm{d}(.,.,.|.) \in \Delta_{\hat{\mathcal{S}}^{\hat{T}-t+1} \times \hat{\mathcal{E}}^{\hat{T}+1} \times \Delta_\Theta | \hat{\mathcal{A}}^{\hat{T}-t+1}}$ as a generic conditional probability distribution. The generic symbol d is used since it represents both Bayesian complete posteriors and approximate complete posteriors. In fact, the definitions of the intrinsic motivations are agnostic with respect to the method used to obtain a complete posterior. In the present context, it is important that these definitions are general enough to induce both Bayesian and variational action-value functions. We usually state the definition of the motivation function using general expressions (e.g., marginalizations) derived from $\mathrm{d}(.,.,.|.)$. Also, we look at how they can be obtained from Bayesian complete posteriors to give to the reader an intuition for the computations involved in applications. The approximate complete posterior usually makes these calculations easier and we will present an example of this.

7.3.1. Free Energy Principle

Here, we present the non-variational Bayesian inference versions for the expressions that occur in the "expected free energy" in Friston et al. (2015, 2017a). These papers only include approximate expressions after variational inference. Most of the expressions we give here can be found in Friston et al. (2017b). The exception is Equation (74), which can be obtained from an approximate term in Friston et al. (2017a) in the same way that the non-variational Bayesian inference terms in Friston et al. (2017b) are obtained from the approximate ones in Friston et al. (2015).

In the following, we can set $\hat{T}_a = \hat{T}$, since actions are only evaluated with respect to their immediate effects.

According to Friston et al. (2017b, Equation (A2) Appendix), the "expected free energy" is just the future conditional entropy of sensor values[6] given environment states. Formally, this is (with a negative sign to make minimizing expected free energy equivalent to maximizing the action-value function):

$$\mathfrak{M}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) := \sum_{\hat{e}_{t:\hat{T}}} \mathrm{d}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}) \sum_{\hat{s}_{t:\hat{T}}} \mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}) \log \mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}) \tag{63}$$

$$= -\sum_{\hat{e}_{t:\hat{T}}} \mathrm{d}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}})\, \mathrm{H}_{\mathrm{d}}(\hat{S}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}) \tag{64}$$

$$= -\mathrm{H}_{\mathrm{d}}(\hat{S}_{t:\hat{T}}|\hat{E}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}). \tag{65}$$

Note that, we indicate the probability distribution d used to calculate entropies $\mathrm{H}_{\mathrm{d}}(X)$ or mutual informations $\mathrm{I}_{\mathrm{d}}(X:Y)$ in the subscript. Furthermore,we indicate the variables that are summed over with capital letters and those that are fixed (e.g., $\hat{a}_{t:\hat{T}}$ above) with small capital letters.

In the case where $\mathrm{d}(.,.,.|.)$ is the Bayesian complete posterior $\mathrm{q}(.,.,.|., sa_{\prec t}, \xi)$, it uses the predictive distribution of environment states $\mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ and the posterior of the conditional distribution of sensor values given environment states $\mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, sa_{\prec t}, \xi)$. As we see next, both distributions can be obtained from the Bayesian complete posterior.

The former distribution is a familiar expression in hierarchical Bayesian models and corresponds to a posterior predictive distribution or predictive density [cmp. e.g., Bishop, 2011, Equation (3.74)] that can be calculated via:

$$\mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$$

$$= \int \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\prec t}} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)\, \mathrm{d}\theta \tag{66}$$

$$= \int \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\prec t}} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)\, \mathrm{d}\theta \tag{67}$$

$$= \int \sum_{\hat{e}_{t-1}} \mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{q}(\hat{e}_{t-1}, \theta|sa_{\prec t}, \xi)\, \mathrm{d}\theta, \tag{68}$$

where we split the complete posterior into the predictive and posterior factor and then marginalized out environment states $\hat{e}_{\prec t-1}$ since the predictive factor does not depend on them. Note that in practice, this marginalization corresponds to a sum over $|\mathcal{E}|^{t-1}$ terms and therefore has a computational cost that grows exponential in time. However, if we use the approximate complete posterior such that $\mathrm{d}(.,.,.|.) = \mathrm{r}(.,.,.|., \phi)$, we see from Equation (40), that $\mathrm{q}(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)$ is replaced by $\mathrm{r}(\hat{e}_{\prec t}, \theta|\phi)$ which is defined as (Equation 38):

$$\mathrm{r}(\hat{e}_{\prec t}, \theta|\phi) := \prod_{\tau=0}^{t-1} \mathrm{r}(\hat{e}_{\tau}|\phi^{E_{\tau}}) \prod_{i=1}^{3} \mathrm{r}(\theta^i|\phi^i). \tag{69}$$

This means that $\mathrm{r}(\hat{e}_{t-1}, \theta|\phi)$ is just $\mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}})\,\mathrm{r}(\theta|\phi)$, which we obtain directly from the variational inference without any marginalization. If Bayesian inference increases in computational cost exponentially in time, this simplification leads to a significant advantage. This formulation leaves an integral over θ or, more precisely, a triple integral over the three $\theta^1, \theta^2, \theta^3$. However, if the $\mathrm{q}(\theta^i|\xi^i)$ are chosen as conjugate priors to $\mathrm{q}(\hat{s}|\hat{e}, \theta^1), \mathrm{q}(\hat{e}'|\hat{a}', \hat{e}, \theta^2), \mathrm{q}(\hat{e}_0|\theta^3)$ respectively, then these integrals can be calculated analytically [compare the similar calculation of $\mathrm{q}(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)$ in **Appendix A**]. The remaining computational problem is only the sum over all $\hat{e}_{t-1}$.

The latter term (the posterior conditional distribution over sensor values given environment states) can be obtained via

$$\mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, sa_{\prec t}, \xi) = \mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \tag{70}$$

$$= \frac{\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)}{\mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)}. \tag{71}$$

Here, the first equation holds since

$$\hat{S}_{t:\hat{T}} \perp\!\!\!\perp \hat{A}_{t:\hat{T}} \mid \hat{E}_{t:\hat{T}}, SA_{\prec t}. \tag{72}$$

Both numerator and denominator can be obtained from the complete posterior via marginalization as for the former term. This marginalization also shows that the intrinsic motivation function, Equation (63), is a functional of the complete posteriors or $\mathrm{d}(.,.,.|.)$.

In most publications on active inference the expected free energy in Equation (63) is only part of what is referred to as the expected free energy. Usually, there is a second term measuring the relative entropy to an externally specified *prior over future outcomes* (also called "predictive distribution encoding goals" Friston et al. 2015), i.e., a desired probability distribution $\mathrm{p}^d(\hat{s}_{t:\hat{T}})$. The relative entropy term is formally given by:

$$\mathrm{KL}[\mathrm{d}(\hat{S}_{t:\hat{T}}|\hat{a}_{t:\hat{T}})||\,\mathrm{p}^d(\hat{S}^d_{t:\hat{T}})] = \sum_{\hat{s}_{t:\hat{T}}} \mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}) \log \frac{\mathrm{d}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}})}{\mathrm{p}^d(\hat{s}_{t:\hat{T}})}. \tag{73}$$

Clearly, this term will lead the agent to act such that the future distribution over sensor values is similar to the desired distribution. Since this term is used to encode extrinsic value for the agent, we mostly ignore it in this publication. It could included into any of the following intrinsic motivations.

In Friston et al. (2017a) yet another term, called "negative novelty" or "ignorance", occurs in the expected free energy. This term concerns the posterior distribution over parameter θ^1. It can be slightly generalized to refer to any subset of the parameters $\theta = (\theta^1, \theta^2, \theta^3)$. We can write it as a conditional

[6] The original text refers to this as the "expected entropy of outcomes," not the expected conditional entropy of outcomes. Nonetheless, the associated Equation (A2) in the original is identical to ours.

mutual information between future sensor values and parameters (the “ignorance” is the negative of this):

$$\mathrm{I_d}(\hat{S}_{t:\hat{T}} : \Theta | \hat{a}_{t:\hat{T}}) = \sum_{\hat{s}_{t:\hat{T}}} \mathrm{d}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}) \int \mathrm{d}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}) \times \log \frac{\mathrm{d}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}})}{\mathrm{d}(\theta)} \, \mathrm{d}\theta. \tag{74}$$

This is identical to the information gain used in knowledge seeking agents. The necessary posteriors in the Bayesian case are $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, $\mathrm{q}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ and $\mathrm{q}(\theta | sa_{\prec t}, \xi)$ with

$$\mathrm{q}(\hat{s}_{t\,:\,\hat{T}} | \hat{a}_{t\,:\,\hat{T}}, sa_{\prec t}, \xi) = \int \sum_{\hat{e}_{\prec t}} \mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \, \mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) \, \mathrm{d}\theta \tag{75}$$

a straightforward (if costly) marginalization of the complete posterior. Just like previously for $\mathrm{q}(\hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, the marginalization is greatly simplified in the variational case (see **Appendix B** for a more explicit calculation). The integrals can be computed if using conjugate priors. The other two posteriors can be obtained via

$$\mathrm{q}(\theta | \hat{s}_{t:\hat{T}}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) = \frac{1}{\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)} \sum_{\hat{e}_{0:\hat{T}}} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi). \tag{76}$$

and

$$\mathrm{q}(\theta | sa_{\prec t}, \xi) = \mathrm{q}(\theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) \tag{77}$$

$$= \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}} \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi). \tag{78}$$

In the latter equation we used

$$\hat{A}_{t:\hat{T}} \perp\!\!\!\perp \Theta | SA_{\prec t}. \tag{79}$$

The marginalizations grow exponentially in computational cost with $\hat{T}$. In this case, the variational approximation only reduces the necessary marginalization over $\hat{e}_{\prec t-1}$ to one over $\hat{e}_{t-1}$, but the marginalization over future environment states $\hat{e}_{t:\hat{T}}$ and sensor values $\hat{s}_{t:\hat{T}}$ remains the same since we use the exact predictive factor. In practice the time horizon into the future $\hat{T} - t$ must then be chosen sufficiently short, so that marginalizing out $\hat{e}_{t:\hat{T}}$ and $\hat{S}_{t:\hat{T}}$ is feasible. Together with the variational approximation the required marginalizations over past and future are then constant over time which makes the implementation of agents with extended lifetimes possible.

The combination of the conditional entropy term and the information gain defines the (intrinsic part) of the action-value function of Friston’s active inference (or free energy principle):

$$\mathfrak{M}^{FEP}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) = -\,\mathrm{H_d}(\hat{S}_{t:\hat{T}} | \hat{E}_{t:\hat{T}}) + \mathrm{I_d}(\hat{S}_{t:\hat{T}} : \theta | \hat{a}_{t:\hat{T}}) \tag{80}$$

In the active inference literature this is usually approximated by a sum over the values at individual timesteps:

$$\mathfrak{M}^{FEP}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) = \sum_{\tau=t}^{\hat{T}} -\,\mathrm{H_d}(\hat{S}_\tau | \hat{E}_\tau) + \mathrm{I_d}(\hat{S}_\tau : \Theta | \hat{a}_{t:\hat{T}}). \tag{81}$$

7.3.2. Free Energy Principle Specialized to Friston et al. (2015)

Using **Appendix C**, we show how to get the action-value function of Friston et al. (2015, Equation 9) in our framework. In Friston et al. (2015), the extrinsic value term of Equation (73) is included, but not the information gain term of Equation (74). Furthermore, the sum over timesteps in Equation (81) is used. This leads to the following expression:

$$\mathfrak{M}^{FEP}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) = \sum_{\tau=t}^{\hat{T}} -\,\mathrm{H_d}(\hat{S}_\tau | \hat{E}_\tau) - \mathrm{KL}[\mathrm{d}(\hat{S}_\tau | \hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)]. \tag{82}$$

If we plug in an approximate complete posterior, we get:

$$\mathfrak{M}^{FEP}(\mathrm{r}(.,.,.|.), \hat{a}_{t:\hat{T}}) = \sum_{\tau=t}^{\hat{T}} -\,\mathrm{H_r}(\hat{S}_\tau | \hat{E}_\tau) - \mathrm{KL}[\mathrm{r}(\hat{S}_\tau | \hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)]. \tag{83}$$

with

$$-\,\mathrm{H_r}(\hat{S}_\tau | \hat{E}_\tau) = \sum_{\hat{e}_\tau} \mathrm{r}(\hat{e}_\tau | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) \sum_{\hat{s}_\tau} \mathrm{r}(\hat{s}_\tau | \hat{e}_\tau, \phi) \log \mathrm{r}(\hat{s}_\tau | \hat{e}_\tau, \phi), \tag{84}$$

and

$$\mathrm{KL}[\mathrm{r}(\hat{S}_\tau | \hat{a}_{t:\hat{T}}) || \, \mathrm{p}^d(\hat{S}_\tau)] = \sum_{\hat{s}_\tau} \mathrm{r}(\hat{s}_\tau | \hat{a}_{t:\hat{T}}, \phi) \log \frac{\mathrm{r}(\hat{s}_\tau | \hat{a}_{t:\hat{T}}, \phi)}{\mathrm{p}^d(\hat{s}_\tau)}. \tag{85}$$

For the particular approximate posterior of Equation (40), with its factorization into exact predictive and approximate posterior factor, the individual terms can be further rewritten.

$$\mathrm{r}(\hat{e}_\tau | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) = \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\tau+1\,:\,\hat{T}} \hat{e}_{t\,:\,\tau-1} \hat{e}_{0:T-2}} \int \mathrm{r}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, \phi) \, \mathrm{d}\theta \tag{86}$$

$$= \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\tau+1\,:\,\hat{T}} \hat{e}_{t\,:\,\tau-1} \hat{e}_{0:T-2}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \times \mathrm{r}(\hat{e}_{\prec t}, \theta | \phi) \, \mathrm{d}\theta \tag{87}$$

$$= \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{\tau+1\,:\,\hat{T}} \hat{e}_{t\,:\,\tau-1} \hat{e}_{0:T-2}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \times \prod_{r=0}^{t-1} \mathrm{r}(\hat{e}_r | \phi^{E_r}) \prod_{i=1}^{3} \mathrm{r}(\theta^i | \phi^i) \, \mathrm{d}\theta \tag{88}$$

$$= \sum_{\hat{e}_{t\,:\,\tau-1}} \int \mathrm{q}(\hat{e}_{t\,:\,\tau-1}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta^2) \times \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}})\, \mathrm{r}(\theta^2|\phi^2)\, \mathrm{d}\theta^2 \tag{89}$$

$$= \left(\sum_{\hat{e}_{t\,:\,\tau-1}} \int \prod_{r=t}^{\tau} \mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}, \theta^2)\, \mathrm{r}(\theta^2|\phi^2)\, \mathrm{d}\theta^2 \right) \times \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}). \tag{90}$$

In Friston et al. (2015), the environment dynamics $\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}, \theta^2)$ are not inferred and are therefore not parameterized:

$$\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}, \theta^2) = \mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}) \tag{91}$$

and are set to the physical environment dynamics:

$$\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}) = \mathrm{p}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}). \tag{92}$$

This means the integral over θ^2 above is trivial and we get:

$$\mathrm{r}(\hat{e}_\tau|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) = \sum_{\hat{e}_{t\,:\,\tau-1}} \prod_{r=t}^{\tau} \mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1})\, \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) \tag{93}$$

In the notation of Friston et al. (2015) (see **Appendix C** for a translation table), we have

$$\mathrm{q}(\hat{e}_r|\hat{a}_r, \hat{e}_{r-1}) = \mathbf{B}(\hat{a}_r)_{\hat{e}_r \hat{e}_{r-1}} \tag{94}$$

where $\mathbf{B}(\hat{a}_r)$ is a matrix, and

$$\mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) = (\hat{s}_{t-1})_{\hat{e}_{t-1}} \tag{95}$$

where $(\hat{s}_{t-1})$ is a vector, so that

$$\mathrm{r}(\hat{e}_\tau|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi) = (\mathbf{B}(\hat{a}_\tau) \cdots \mathbf{B}(\hat{a}_t) \cdot \hat{s}_{t-1})_{\hat{e}_\tau} \tag{96}$$

$$=: (\hat{s}_\tau(\hat{a}_{t:\hat{T}}))_{\hat{e}_\tau} \tag{97}$$

Similarly, since the sensor dynamics in Friston et al. (2015) are also not inferred, we find

$$\mathrm{r}(\hat{s}_\tau|\hat{e}_\tau, \phi) = \mathrm{q}(\hat{s}_\tau|\hat{e}_\tau) = \mathrm{p}(\hat{s}_\tau|\hat{e}_\tau). \tag{98}$$

Friston et al. writes:

$$\mathrm{q}(\hat{s}_\tau|\hat{e}_\tau) =: \mathbf{A}_{\hat{s}_\tau \hat{e}_\tau} \tag{99}$$

with **A** a matrix. So that,

$$\mathrm{r}(\hat{s}_\tau|\hat{a}_{t:\hat{T}}, \phi^{E_{t-1}}) = \mathbf{A} \cdot \hat{s}_\tau(\hat{a}_{t:\hat{T}}) \tag{100}$$

$$=: \hat{o}_\tau(\hat{a}_{t:\hat{T}}). \tag{101}$$

Then

$$\mathrm{H}_\mathrm{r}(\hat{S}_\tau|\hat{E}_\tau) = -\mathbf{1} \cdot (\mathbf{A} \times \log \mathbf{A}) \cdot \hat{s}_\tau(\hat{a}_{t:\hat{T}}) \tag{102}$$

where × is a Hadamard product and **1** is a vector of ones. Also,

$$\mathrm{KL}[\mathrm{r}(\hat{S}_\tau|\hat{a}_{t:\hat{T}})||\, \mathrm{p}^d(\hat{S}_\tau)] = \hat{o}_\tau(\hat{a}_{t:\hat{T}}) \cdot (\log \hat{o}_\tau(\hat{a}_{t:\hat{T}}) - \log \mathbf{C}_\tau) \tag{103}$$

where $(\mathbf{C}_\tau)_{\hat{s}_\tau} = \mathrm{p}^d(\hat{s}_\tau)$. Plugging these expressions into Equation (83), substituting $\hat{a}_{t:\hat{T}} \to \pi$, and comparing this to Friston et al. (2015, Equation 9) shows that[7]:

$$\mathfrak{M}^{FEP}(\mathrm{r}(.,.,.|.), \pi) = \mathbf{1} \cdot (\mathbf{A} \times \log \mathbf{A}) \cdot \hat{s}_\tau(\hat{a}_{t:\hat{T}}) - \hat{o}_\tau(\hat{a}_{t:\hat{T}}) \cdot (\log \hat{o}_\tau(\hat{a}_{t:\hat{T}}) - \log \mathbf{C}_\tau) \tag{104}$$

$$= \mathbf{Q}(\pi). \tag{105}$$

This verifies that our formulation of the action-value function specializes to the "expected (negative) free energy" $\mathbf{Q}(\pi)$.

7.3.3. Empowerment Maximization

Empowerment maximization (Klyubin et al., 2005) is an intrinsic motivation that seeks to maximize the channel capacity from sequences of the agent's actions into the subsequent sensor value. The agent, equipped with complete knowledge of the environment dynamics, can directly observe the environment state. If the environment is deterministic, an empowerment maximization policy leads the agent to a state from which it can reach the highest number of future states within a preset number of actions.

Salge et al. (2014) provide a good overview of existing research on empowerment maximization. A more recent study relates the intrinsic motivation to the essential dynamics of living systems, based on assumptions from autopoietic enactivism Guckelsberger and Salge (2016). Several approximations have been proposed, along with experimental evaluations in complex state / action spaces. Salge et al. (2018) show how deterministic empowerment maximization in a three-dimensional grid-world can be made more efficient by different modifications of UCT tree search. Three recent studies approximate stochastic empowerment and its maximization via variational inference and deep neural networks, leveraging a variational bound on the mutual information proposed by Barber and Agakov (2003). Mohamed and Rezende (2015) focus on a model-free approximation of open-loop empowerment, and Gregor et al. (2016) propose two means to approximate closed-loop empowerment. While these two approaches consider both applications in discrete and continuous state / action spaces, Karl et al. (2017) develop an open-loop, model-based approximation for the continuous domain specifically. The latter study also demonstrates how empowerment can yield good performance in established reinforcement learning benchmarks such as bipedal balancing in the absence of extrinsic rewards. In recent years, research on empowerment has particularly focused on applications in multi-agent systems. Coupled empowerment maximization as a specific multi-agent policy has been proposed as intrinsic drive for either supportive or antagonistic behaviour in open-ended scenarios with sparse reward landscapes Guckelsberger et al. (2016b). This theoretical investigation has then been backed up with empirical evaluations on supportive and adversarial video game characters Guckelsberger et al. (2016a, 2018). Beyond virtual agents, the same policy has been proposed as a

[7] There is a small typo in Friston et al. (2015, Equation 9) where the time index of $\hat{s}_{t-1}$ in $(\hat{s}_\tau(\hat{a}_{t:\hat{T}})) = (\mathbf{B}(\hat{a}_\tau) \cdots \mathbf{B}(\hat{a}_t) \cdot \hat{s}_{t-1})$ is given as t instead of $t-1$.

good heuristic to facilitate critical aspects of human-robot interaction, such as self-preservation, protection of the human partner, and response to human actions Salge and Polani (2017).

For empowerment, we select $\hat{T}_a = t + n$ and $\hat{T} = t + n + m$, with $n \geq 0$ and $m \geq 1$. This means the agent chooses $n+1$ actions which it expects to maximize the resulting m-step empowerment. The according action-value function is:

$$\mathfrak{M}^{EM}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}_a}) := \max_{\mathrm{d}(\hat{a}_{\hat{T}_a+1:\hat{T}})} \mathrm{I}_{\mathrm{d}}(\hat{A}_{\hat{T}_a+1:\hat{T}} : \hat{S}_{\hat{T}} | \hat{a}_{t:\hat{T}_a}) \quad (106)$$

$$= \max_{\mathrm{d}(\hat{a}_{\hat{T}_a+1:\hat{T}})} \sum_{\hat{a}_{\hat{T}_a+1:\hat{T}}, \hat{s}_{\hat{T}}} \mathrm{d}(\hat{a}_{\hat{T}_a+1:\hat{T}}) \times \mathrm{d}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}}) \log \frac{\mathrm{d}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}})}{\mathrm{d}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}_a})}. \quad (107)$$

Note that in the denominator of the fraction, the action sequence only runs to $t{:}\hat{T}_a$ and not to $t{:}\hat{T}$ as in the numerator.

In the Bayesian case, the required posteriors are $\mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ (for each $\hat{a}_{\hat{T}_a+1:\hat{T}}$) and $\mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi)$. The former distribution is a further marginalization over $\hat{s}_{t+1:\hat{T}-1}$ of $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The variational approximation only helps getting $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, not the further marginalization. The latter distribution is obtained for a given $\mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}})$ from the former one via

$$\mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi) = \sum_{\hat{a}_{\hat{T}_a+1:\hat{T}}} \mathrm{q}(\hat{s}_{\hat{T}}, \hat{a}_{\hat{T}_a+1:\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi) \quad (108)$$

$$= \sum_{\hat{a}_{\hat{T}_a+1:\hat{T}}} \mathrm{q}(\hat{s}_{\hat{T}} | \hat{a}_{\hat{T}_a+1:\hat{T}}, \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi)\, \mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}}) \quad (109)$$

since the empowerment calculation imposes

$$\mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}} | \hat{a}_{t:\hat{T}_a}, sa_{\prec t}, \xi) = \mathrm{q}(\hat{a}_{\hat{T}_a+1:\hat{T}}). \quad (110)$$

7.3.4. Predictive Information Maximization

Predictive information maximization, (Ay et al., 2008), is an intrinsic motivation that seeks to maximize the predictive information of the sensor process. Predictive information is the mutual information between past and future sensory signal, and has been proposed as a general measure of complexity of stochastic processes (Bialek and Tishby, 1999). For applications in the literature see Ay et al. (2012); Martius et al. (2013, 2014). Also, see Little and Sommer (2013) for a comparison to entropy minimization.

For predictive information, we select a half time horizon $k = \lfloor (t{:}\hat{T} - t + 1)/2 \rfloor$ where $k > 0$ for predictive information to be defined (i.e., $t{:}\hat{T} - t > 0$). Then, we can define the expected mutual information between the next m sensor values and the subsequent m sensor values as the action-value function of predictive information maximization. This is similar to the time-local predictive information in Martius et al. (2013):

$$\mathfrak{M}^{PI}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) := \mathrm{I}_{\mathrm{d}}(\hat{S}_{t:t+k-1} : \hat{S}_{t+k:t+2k-1} | \hat{a}_{t:\hat{T}}). \quad (111)$$

We omit writing out the conditional mutual information since it is defined in the usual way. Note that it is possible that $t + 2k - 1 < t{:}\hat{T}$ so that the action sequence $\hat{a}_{t:\hat{T}}$ might go beyond the evaluated sensor probabilities. This displacement leads to no problem since the sensor values do not depend on future actions. The posteriors needed are: $\mathrm{q}(\hat{s}_{t:t+k-1} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, $\mathrm{q}(\hat{s}_{t+k:t+2k-1} | \hat{s}_{t:t+k-1}, \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$, and $\mathrm{q}(\hat{s}_{t+k:t+2k-1} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The first and the last are again marginalizations of $\mathrm{q}(\hat{s}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ seen in Equation (75). The second posterior is a fraction of such marginalizations.

7.3.5. Knowledge Seeking

Knowledge seeking agents (Storck et al., 1995; Orseau et al., 2013) maximize the information gain with respect to a probability distribution over environments. The information gain we use here is the relative entropy between the belief over environments after actions and subsequent sensor values and the belief over environments (this is the KL-KSA of Orseau et al. 2013, "KL" for Kullback-Leibler divergence). In our case the belief over environments can be identified with the posterior $\mathrm{q}(\theta | sa_{\prec t}, \xi)$ since every $\theta = (\theta^1, \theta^2, \theta^3)$ defines an environment. In principle, this can be extended to the posterior $\mathrm{q}(\xi | sa_{\prec t}, \xi)$ over the hyperprior ξ, but we focus on θ here. This definition is more similar to the original one. Then, we define the knowledge seeking action-value function using the information gain of Equation (74):

$$\mathfrak{M}^{KSA}(\mathrm{d}(.,.,.|.), \hat{a}_{t:\hat{T}}) := \mathrm{I}_{\mathrm{d}}(\hat{S}_{t:\hat{T}} : \Theta | \hat{a}_{t:\hat{T}}). \quad (112)$$

We have discussed the necessary posteriors following Equation (74).

After this overview of some intrinsic motivations, we look at active inference. However, what should be clear is, that, in principle, both the posteriors needed for the intrinsic motivation function of the original active inference (Friston et al., 2015) and the posteriors needed for alternative inferences overlap. This overlap shows that the other intrinsic motivations mentioned here also profit from variational inference approximations. There is also no indication that these intrinsic motivations cannot be used together with the next discussed active inference.

8. ACTIVE INFERENCE

Now, we look at active inference. Note that this section is independent of the intrinsic motivation function underlying the action-value function $\hat{\mathrm{Q}}$.

In the following we first look at and try to explain a slightly simplified version of the active inference in Friston et al. (2015). Afterwards we also state the full version.

As mentioned in the introduction, current active inference versions are formulated as an optimization procedure that, at least at first sight, looks similar to the optimization of a variational free energy familiar from variational inference. Recall that, in variational inference the parameters of a family of distributions are optimized to approximate an exact (Bayesian) posterior of a generative model. In the case we discussed in Section 6.4 the sought after exact posterior is the posterior factor of the generative model of Section 6.1. One of our questions about active inference is whether it is a straightforward application of variational inference to a posterior of some generative model. This would imply the existence of a generative model whose standard updating with past actions and sensor values leads to an optimal posterior distribution over future actions. Note that, this does not work with the generative model in of Section 6.1 since the future actions there are independent of the past sensor values and actions. Given the appropriate generative model, it would then be natural to introduce it first and then apply a variational approximation similar to our procedure in Section 6.

We were not able to find in the literature or construct ourselves a generative model such that variational inference leads directly to the active inference as given in Friston et al. (2015). Instead we present a generative model that contains a posterior whose variational approximation optimization is very similar to the optimization procedure of active inference. It is also closely related to the two-step action generation of first inferring the posterior and then selecting the optimal actions. This background provides some intuition for the particularities of active inference.

One difference of the generative model used here is that its structure depends on the current time step in a systematic way. The previous generative model of Section 6.1 had a time-invariant structure.

In Section 6, we showed how the generative model, together with either Bayesian or variational inference, can provide an agent with a set of complete posteriors. Each complete posterior is a conditional probability distribution over all currently unobserved variables ($\hat{S}_{t:\hat{T}}, \hat{E}_{0:T}$) and parameters ($\Theta$ and more generally also Ξ) given past sensor values and actions $sa_{\prec t}$ and a particular sequence of future actions $\hat{a}_{t:\hat{T}}$. Inference means updating the set of posteriors in response to observations $sa_{\prec t}$. Active inference should then update the distribution over future actions in response to observations. This means the according posterior cannot be conditional on future action sequences like the complete posterior in Equation (16). Since active inference promises belief or knowledge updating and action selection in one mechanism the posterior should also range over unobserved relevant variables like future sensor values, environment states, and parameters. This leads to the posterior of Equation (13):

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi). \tag{13 revisited}$$

If this posterior has the right structure, then we can derive a future action distribution by marginalizing:

$$\mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) = \sum_{\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi)\, \mathrm{d}\theta. \tag{113}$$

Actions can then be sampled from the distribution obtained by marginalizing further to the next action only:

$$\mathrm{p}(a_t | m_t) := \sum_{\hat{a}_{t+1\,:\,\hat{T}}} \mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi). \tag{114}$$

This scheme could justifiably be called (non-variational) active inference since the future action distribution is directly obtained by updating the generative model.

However, as we mentioned above, according to the generative model of **Figure 2**, the distribution over future actions is independent of the past sensor values and actions:

$$\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi) = \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)\, \mathrm{q}(\hat{a}_{t:\hat{T}}) \tag{115}$$

since

$$\mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) = \mathrm{q}(\hat{a}_{t:\hat{T}}). \tag{116}$$

Therefore, we can never learn anything about future actions from past sensor values and actions using this model. In other words, if we intend to select the actions based on the past, we cannot uphold this independent model. The inferred actions must become dependent on the history and the generative model has to be changed for a scheme like the one sketched above to be successful.

In Section 7.2, we have mentioned that the softmax policy based on a given action-value function $\hat{Q}$ could be a desirable outcome of an active inference scheme such as the above. Thus, if we ended up with

$$\mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi) = \frac{1}{Z(\gamma, sa_{\prec t}, \xi)} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)} \tag{117}$$

as a result of some active inference process, that would be a viable solution. We can force this by building this conditional distribution directly into a new generative model. Note that this conditional distribution determines all future actions $\hat{a}_{t:\hat{T}}$ starting at time t and not just the next action $\hat{a}_t$. In the end however only the next action will be taken according to Equation (114) and at time $t+1$ the action generation mechanism starts again, now with $\hat{a}_{t+1\,:\,\hat{T}}$ influenced by the new data sa_t in addition to $sa_{\prec t}$. So the model structure changes over time in this case with the dependency of actions on pasts $sa_{\prec t}$ shifting together with each time-step. Keeping the rest of the previous Bayesian network structure intact we define that at each time t the next action $\hat{A}_t$ depends on past sensor values and actions $sa_{\prec t}$ as well as on the hyperparameter ξ (see **Figure 6**):

$$\begin{aligned}\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \theta | sa_{\prec t}, \xi) :=\ & \mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}} | \hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \\ & \times\ \mathrm{q}(\hat{a}_{t:\hat{T}} | sa_{\prec t}, \xi)\, \mathrm{q}(\theta, \hat{e}_{\prec t} | sa_{\prec t}, \xi).\end{aligned} \tag{118}$$

On the right hand side we have the predictive and posterior factors left and right of the distribution over future actions. We define this conditional future action distribution to be the softmax of Equation (117). This means that the mechanism-generating future actions uses the Bayesian action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$. The Bayesian action-value function depends on the complete posterior $\mathrm{q}(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}, \theta | \hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ calculated using the old generative model of **Figure 2** where actions do

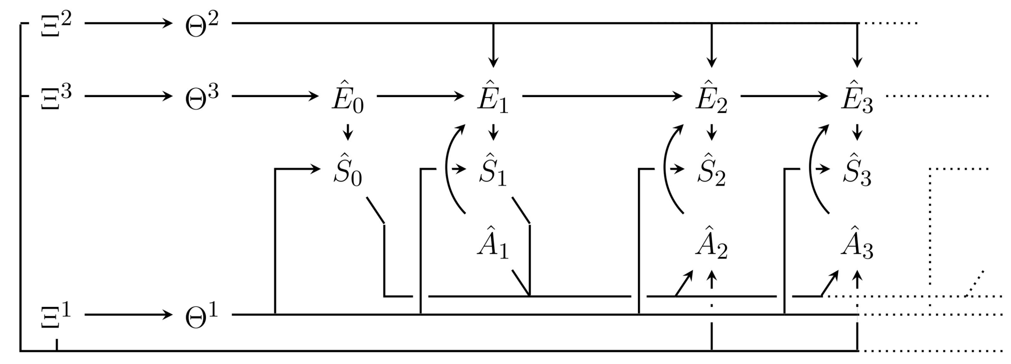


FIGURE 6 | Generative model including $q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)$ at $t = 2$ with $\hat{S}\hat{A}_{\prec 2}$ influencing future actions $\hat{A}_{2:\hat{T}}$. Note that, only future actions are dependent on past sensor values and actions, e.g., action $\hat{A}_1$ has no incoming edges. The increased gap between time step $t = 1$ and $t = 2$ is to indicate that this time step is special in the model. For each time step t there is an according model with the particular relation between past $\hat{S}\hat{A}_{\prec t}$ and $\hat{A}_{t:\hat{T}}$ shifted accordingly.

not not depend on past sensor values and actions. This is a complex construction with what amounts to Bayesian inference essentially happening within an edge (i.e., $\hat{S}\hat{A}_{\prec t} \rightarrow \hat{A}_{t:\hat{T}}$) of a Bayesian network. However, logically there is no problem since the posterior $q(\hat{s}_{t:\hat{T}}, \hat{e}_{t:\hat{T}}, \theta|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ for each $\hat{a}_{t:\hat{T}}$ to be well defined really only needs $sa_{\prec t}$, ξ, and the model structure. Here we see the model structure as "hard wired" into the mechanism, since it is fixed for each time step t from the beginning.

We now approximate the posterior of Equation (117) using variational inference. Like in Section 6.4 we do not approximate the predictive factor. Instead we only approximate the product of posterior factor $q(\theta, \hat{e}_{\prec t}|sa_{\prec t}, \xi)$ and future action distribution $q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)$. By construction these are two independent factors but with an eye to active inference which treats belief or knowledge updating and action generation together we also treat them together. For the approximation we again use the approximate posterio factor of Equation (38) and combine it with a distribution over future actions $r(\hat{a}_{t:\hat{T}}|\pi)$ parameterized by π:

$$r(\hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}, \theta|\pi, \phi) := r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi) \tag{119}$$

$$:= r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}|\phi^{E_{\prec t}})\, r(\theta|\phi). \tag{120}$$

The variational free energy is then:

$$\mathcal{F}[\pi, \phi, sa_{\prec t}, \xi] := \sum_{\hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}} \int r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi) \times \log \frac{r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi)}{q(s_{\prec t}, \hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}, \theta|a_{\prec t}, \xi)}\, d\theta \tag{121}$$

$$= \sum_{\hat{a}_{t:\hat{T}}, \hat{e}_{\prec t}} \int r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi) \times \log \frac{r(\hat{a}_{t:\hat{T}}|\pi)\, r(\hat{e}_{\prec t}, \theta|\phi)}{q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)\, q(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)\, q(s_{\prec t}|a_{\prec t}, \xi)}\, d\theta \tag{122}$$

$$= \mathcal{F}[\phi, sa_{\prec t}, \xi] + \mathrm{KL}[r(\hat{A}_{t:\hat{T}}|\pi)||\, q(\hat{A}_{t:\hat{T}}|sa_{\prec t}, \xi)]. \tag{123}$$

Where $\mathcal{F}[\phi, sa_{\prec t}, \xi]$ is the variational free energy of the (non-active) variational inference (see Equation 45). Variational inference then minimizes the above expression with respect to parameters ϕ and π:

$$\phi^*_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} := \arg\min_{\phi,\pi} \mathcal{F}[\pi, \phi, sa_{\prec t}, \xi]$$

$$= \arg\min_{\phi} \mathcal{F}[\phi, sa_{\prec t}, \xi] \tag{124}$$

$$+ \arg\min_{\pi} \mathrm{KL}[r(\hat{A}_{t:\hat{T}}|\pi)||\, q(\hat{A}_{t:\hat{T}}|sa_{\prec t}, \xi)]. \tag{125}$$

We see that the minimization in this case separates into two minimization problems. The first is just the variational inference of Section 6.4 and the second minimizes the KL-divergence between the parameterized action distribution $r(\hat{a}_{t:\hat{T}}|\pi)$ and the softmax $q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi)$ of the Bayesian action-value function. It is instructive to look at this KL-divergence term closer:

$$\mathrm{KL}[r(\hat{A}_{t:\hat{T}}|\pi)||\, q(\hat{A}_{t:\hat{T}}|sa_{\prec t}, \xi)] = -\mathrm{H}_r(\hat{A}_{t:\hat{T}}|\pi) - \sum_{\hat{a}_{t:\hat{T}}} r(\hat{a}_{t:\hat{T}}|\pi) \log q(\hat{a}_{t:\hat{T}}|sa_{\prec t}, \xi) \tag{126}$$

$$= -\mathrm{H}_r(\hat{A}_{t:\hat{T}}|\pi) - \sum_{\hat{a}_{t:\hat{T}}} r(\hat{a}_{t:\hat{T}}|\pi) \hat{Q}(\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi) + \log Z(\gamma, sa_{\prec t}, \xi). \tag{127}$$

We see that the optimization of π leads toward high entropy distributions for which the expectation value of the action-value function $\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$ is large. Action selection could then happen according to

$$p(a_t|m_t) := \sum_{\hat{a}_{t+1:T}} r(\hat{a}_{t:\hat{T}}|\pi^*_{sa_{\prec t},\xi}). \tag{128}$$

So the described variational inference procedure, at least formally, leads to a useful result. However, this is not the active inference procedure of Friston et al. (2015). As noted above the minimization actually splits into two completely independent minimizations here. The result of the minimization with respect to ϕ in Equation (125) is actually not used for action selection and since action selection is all that matters here is mere ornament. However, there is a way to make use of it. Recall that plugging $\phi^*_{sa_{\prec t},\xi}$ into the variational action-value function $\hat{Q}(\hat{a}_{t:\hat{T}},\phi)$ means that it approximates the Bayesian action value function (see Equation 52). This means that if we define a softmax distribution $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$ of the variational action-value function parameterized by ϕ as:

$$\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi) = \frac{1}{Z(\gamma,\phi)} e^{\gamma \hat{Q}(\hat{a}_{t:\hat{T}},\phi)}. \tag{129}$$

Then this approximates the softmax of the Bayesian action-value function:

$$\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi^*_{sa_{\prec t},\xi}) \approx \mathrm{q}(\hat{a}_{t:\hat{T}}|sa_{\prec t},\xi). \tag{130}$$

Consequently, once we have obtained $\phi^*_{sa_{\prec t},\xi}$ from the first minimization problem in Equation (125) we can plug it into $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$ and then minimize the KL-divergence of $\mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)$ to this distribution instead of the one to $\mathrm{q}(\hat{a}_{t:\hat{T}}|sa_{\prec t},\xi)$. In this way the result of the first could be reused for the second minimization. This remains a two part action generation mechanism however. Active inference combines these two steps into one minimization by replacing $\mathrm{q}(\hat{a}_{t:\hat{T}}|sa_{\prec t},\xi)$ in the variational free energy of Equation (121) with $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$. Since $\mathrm{r}(\hat{a}_{t:\hat{T}}|\phi)$ thereby becomes part of the denominator it is also given the same symbol (in our case q) as the generative model. So we define:

$$\mathrm{q}(\hat{a}_{t:\hat{T}}|\phi) := \mathrm{r}(\hat{a}_{t:\hat{T}}|\phi). \tag{131}$$

In this form the softmax $\mathrm{q}(\hat{a}_{t:\hat{T}}|\phi)$ is a cornerstone of active inference. In brief, it can be regarded as a prior over action sequences. To obtain purposeful behaviour it specifies prior assumptions about what sorts of actions an agent should take when its belief parameter takes value ϕ. Strictly speaking the expression resulting from the replacement $\mathrm{q}(\hat{A}_{t:\hat{T}}|sa_{\prec t},\xi) \rightarrow \mathrm{q}(\hat{a}_{t:\hat{T}}|\phi)$ in Equation (121) is then not a variational free energy anymore since the variational parameters ϕ occur in both the numerator and the denominator. Nonetheless, this is the functional that is minimized in active inference as described in Friston et al. (2015). So active inference is defined as the optimization problem (cmp. Friston et al., 2015, Equation 1):

$$\begin{aligned}\phi^*_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} &= \underset{\phi,\pi}{\arg\min} \sum_{\hat{a}_{t:\hat{T}},\hat{e}_{\prec t}} \int \mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\,\mathrm{r}(\hat{e}_{\prec t},\theta|\phi) \\ &\qquad \log \frac{\mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\,\mathrm{r}(\hat{e}_{\prec t},\theta|\phi)}{\mathrm{q}(s_{\prec t},\hat{a}_{t:\hat{T}},\hat{e}_{\prec t},\theta|\phi,a_{\prec t},\xi)}\,\mathrm{d}\theta \qquad (132) \\ &= \underset{\phi,\pi}{\arg\min} \, (\mathcal{F}[\phi, sa_{\prec t},\xi] \\ &\qquad + \mathrm{KL}[\mathrm{r}(\hat{A}_{t:\hat{T}}|\pi) || \, \mathrm{q}(\hat{A}_{t:\hat{T}}|\phi)]\Big). \qquad (133)\end{aligned}$$

This minimization does not split into the two independent parts anymore since both the future action distribution $\mathrm{q}(\hat{A}_{t:\hat{T}}|\phi)$ of the generative model and the approximate posterior factor in the variational free energy $\mathcal{F}[\phi, sa_{\prec t},\xi]$ are parameterized by ϕ. This justifies the claim that active inference obtains both belief update and action selection through a single principle or optimization.

Compared to Friston et al. (2015), we have introduced a simplification of active inference. In the original text, additional distributions over γ (with according random variable Γ) are introduced to the generative model as $\mathrm{q}(\gamma|\xi^\Gamma)$ (which is a fixed prior) and to the approximate posterior as $\mathrm{r}(\gamma|\phi^\Gamma)$. For the sake of completeness, we show the full equations as well. Since γ is now part of the model, we write $\mathrm{q}(\hat{a}_{t:\hat{T}}|\gamma,\phi)$ instead of $\mathrm{q}(\hat{a}_{t:\hat{T}}|\phi)$. The basic procedure above stays the same. The active inference optimization becomes:

$$\begin{aligned}&\phi^*_{sa_{\prec t},\xi}, \phi^{\Gamma *}_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} \\ &\quad = \underset{\phi,\phi^\Gamma,\pi}{\arg\min} \sum_{\hat{a}_{t:\hat{T}},\hat{e}_{\prec t}} \iint \mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\,\mathrm{r}(\gamma|\phi^\Gamma)\,\mathrm{r}(\hat{e}_{\prec t},\theta|\phi) \\ &\quad \times \log \frac{\mathrm{r}(\hat{a}_{t:\hat{T}}|\pi)\,\mathrm{r}(\gamma|\phi^\Gamma)\,\mathrm{r}(\hat{e}_{\prec t},\theta|\phi)}{\mathrm{q}(s_{\prec t},\hat{a}_{t:\hat{T}},\gamma,\hat{e}_{\prec t},\theta|\phi,a_{\prec t},\xi)}\,\mathrm{d}\theta\,\mathrm{d}\gamma. \qquad (134)\end{aligned}$$

Note that here, by construction, the denominator can be written as:

$$\begin{aligned}&\mathrm{q}(s_{\prec t},\hat{a}_{t:\hat{T}},\gamma,\hat{e}_{\prec t},\theta|\phi,a_{\prec t},\xi) \\ &\quad = \mathrm{q}(\hat{a}_{t:\hat{T}}|\gamma,\phi)\,\mathrm{q}(\gamma|\phi^\Gamma)\,\mathrm{q}(\hat{e}_{\prec t},\theta|sa_{\prec t},\xi)\,\mathrm{q}(s_{\prec t}|a_{\prec t},\xi). \qquad (135)\end{aligned}$$

Which allows us to write Equation (134) with the original variational free energy again:

$$\begin{aligned}\phi^*_{sa_{\prec t},\xi}, \phi^{\Gamma *}_{sa_{\prec t},\xi}, \pi^*_{sa_{\prec t},\xi} &= \underset{\phi,\phi^\Gamma,\pi}{\arg\min} \, (\mathcal{F}[\phi, sa_{\prec t},\xi] \\ &\qquad + \mathrm{KL}[\mathrm{r}(\hat{A}_{t:\hat{T}},\Gamma|\pi,\phi^\Gamma) || \, \mathrm{q}(\hat{A}_{t:\hat{T}},\Gamma|\phi,\xi^\Gamma)]\Big). \qquad (136)\end{aligned}$$

9. APPLICATIONS AND LIMITATIONS

An application of the active inference described here to a simple maze task can be found in Friston et al. (2015). Active inference using different forms of approximate posteriors can be found in Friston et al. (2016b). Here, Friston et al. (2017a) also includes a knowledge seeking term in addition to the conditional entropy term. In the universal reinforcement learning framework Aslanides et al. (2017) also implement a knowledge seeking agent. These works can be quite directly translated into our framework.

For applications of intrinsic motivations that are not so directly related to our framework see also the references in the according Sections 7.3.3 to 7.3.5.

A quantitative analysis of the limitations of the different approaches we discussed is beyond the scope of this publication. However, we can make a few observations that may help researchers interested in applying the discussed approaches.

Concerning the computation of the complete posterior by direct Bayesian methods is not feasible beyond the simplest of systems and even then only for very short time durations. As mentioned in the text it contains a sum over $|\hat{\mathcal{E}}|^t$ elements. If the time horizon into the future is $\hat{T} - t$ then the predictive factor consists of $\hat{\mathcal{S}}^{\hat{T}-t} \times \hat{\mathcal{E}}^{\hat{T}-t} \times \hat{\mathcal{A}}^{\hat{T}-t}$ entries. This means predicting far into the future is also not feasible. Therefore $\hat{T}-t$ will usually have to be fixed to a small number. Methods that also approximate the predictive factor (e.g., Friston et al., 2016b, 2017a) may be useful here. However, to our knowledge, their scalability has not been addressed yet. Since in these approaches the predictive factor is approximated in a similar way as the posterior factor here, we would expect that it is similar to the scalability of approximating the posterior factor.

Employing variational inference reduces the computational burden for obtaining a posterior factor considerably. The sum over all possible past environment histories (the $|\hat{\mathcal{E}}|^t$ elements) is approximated within the optimization. Clearly, by employing variational inference we inherit all shortcomings of this method. As mentioned also in Friston et al. (2016b) variational inference approximations are known to become overconfident i.e., the approximate posterior tends to ignore values with low probabilities (see e.g., Bishop, 2011). In practice this can of course lead to poor decision making. Furthermore, the convergence of the optimization to obtain the approximate posterior can also become slow. As time t increases the necessary computations for each optimization step in the widely used coordinate ascent variational inference algorithm (Blei et al., 2017) grow with t^2. Experiments suggest that the number of necessary optimization steps also grows over time. At the moment, we do not know how fast but this may also lead to problems. A possible solution would be to introduce some form of forgetting such that the considered past does not grow forever.

Ignoring the problem of obtaining a complete posterior, we still have to evaluate and select actions. Computing the information theoretic quantities needed for the mentioned intrinsic motivations and their induced action-value functions is also computationally expensive. In this case fixing the future time horizon $\hat{T} - t$ can lead to constant computational requirements. These grow exponentially with the time horizon which makes large time horizons impossible without further approximations. Note that the action selection mechanisms discussed here also require the computation of the action-value functions for each of the future action sequences.

Active inference is not a standard variational inference problem and therefore standard algorithms like the coordinate ascent variational inference may fail in this case. Other optimization procedures like gradient descent may still work. As far as we know there have been no studies of the scalability of the active inference scheme up to now.

10. CONCLUSION

We have reconstructed the active inference approach of Friston et al. (2015) in in a formally consistent way. We started by disentangling the components of inference and action selection. This disentanglement has allowed us to also remove the variational inference completely and formulate the pure Bayesian knowledge updating for the generative model of Friston et al. (2015). We have shown in Section 6.3 that a special case of this model is equivalent to a finite version of the model used by the Bayesian universal reinforcement agent (Hutter, 2005). We then pointed out how to approximate the pure Bayesian knowledge updating with variational inference. To formalize the notion of intrinsic motivations within this framework, we have introduced intrinsic motivation functions that take complete posteriors and future actions as inputs. These induce action-value functions similar to those used in reinforcement learning. The action-value functions can then be used for both, the Bayesian and the variational agent, in standard deterministic or softmax action selection schemes.

Our analysis of the intrinsic motivations *Expected Free Energy Maximization*, *Empowerment Maximization*, *Predictive Information Maximization*, and *Knowledge Seeking* indicates that there is significant common structure between the different approaches and it may be possible to combine them. At the time of writing, we have already made first steps toward using the present framework for a systematic quantitative analysis and comparison of the different intrinsic motivations. Eventually, such studies will shed more conclusive light on the computational requirements and emergent dynamics of different motivations. An investigation of the biological plausibility of different motivations might lead to different results and this is of equal interest.

Beyond the comparison of different intrinsic motivations within an active inference framework, the present work can thus contribute to investigations on the role of intrinsic motivations in living organisms. If biological plausibility of active inference can be upheld, and maintained for alternative intrinsic motivations, then experimental studies might be derived to test differentiating predictions. If active inference was key to cognitive phenomena such as consciousness, it would be interesting to see how the cognitive dynamics would be affected by alternative intrinsic motivations.

AUTHOR CONTRIBUTIONS

MB, CG, CS, SS, and DP conceived of this study, discussed the concepts, revised the formal analysis, and wrote the article. MB contributed the initial formal analysis.

FUNDING

CG is funded by EPSRC grant [EP/L015846/1] (IGGI). CS is funded by the EU Horizon 2020 programme under the Marie Sklodowska-Curie grant 705643. DP is funded in part by EC H2020-641321 socSMCs FET Proactive project.

ACKNOWLEDGMENTS

MB would like to thank Yen Yu for valuable discussions on active inference.

REFERENCES

Allen, M., and Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese* 195, 2459–2482. doi: 10.1007/s11229-016-1288-5

Aslanides, J., Leike, J., and Hutter, M. (2017). "Universal reinforcement learning algorithms: survey and experiments," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, VIC), 1403–1410.

Attias, H. (1999). "A variational Bayesian framework for graphical models," in *Proceedings Advances in Neural Information Processing Systems 12*, eds S. Solla, T. Leen, and K. Müller (Cambridge, MA: MIT Press), 209–215.

Attias, H. (2003). "Planning by probabilistic inference," in *Proceedings 9th International Workshop on Artificial Intelligence and Statistics* (Key West, FL).

Ay, N., Bernigau, H., Der, R., and Prokopenko, M. (2012). Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theor. Biosci.* 131, 161–179. doi: 10.1007/s12064-011-0137-9

Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *Eur. Phys. J. B Cond. Matter Complex Syst.* 63, 329–339. doi: 10.1140/epjb/e2008-00175-0

Ay, N. and Löhr, W. (2015). The umwelt of an embodied agent–a measure-theoretic definition. *Theor. Biosci.* 134, 105–116. doi: 10.1007/s12064-015-0217-3

Barber, D., and Agakov, F. (2003). "The IM algorithm: a variational approach to information maximization," in *Proceedings Advances in Neural Information Processing Systems 16*, eds S. Thrun, L. K. Saul, and B. Schölkopf (Vancouver, BC: MIT Press), 201–208.

Bialek, W., and Tishby, N. (1999). Predictive information. *arXiv:cond-mat/9902341*.

Bishop, C. M. (2011). *Pattern Recognition and Machine Learning. Information Science and Statistics*. New York, NY: Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

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

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *J. Math. Psychol.* 81, 55–79. doi: 10.1016/j.jmp.2017.09.004

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, N.J: Wiley-Interscience.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin Books.

Doshi-Velez, F., Pfau, D., Wood, F., and Roy, N. (2015). Bayesian nonparametric methods for partially-Observable reinforcement learning. *IEEE Trans. Patt. Anal. Mach. Intell.* 37, 394–407. doi: 10.1109/TPAMI.2013.191

Ellis, B., and Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. *J. Am. Stat. Assoc.* 103, 778–789. doi: 10.1198/016214508000000193

Fox, R., and Tishby, N. (2016). "Minimum-information LGQ control part II: retentive controllers," in *2016 IEEE 55th Conference on Decision and Control (CDC)* (Las Vegas), 5603–5609.

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

Friston, K. (2013a). Consciousness and hierarchical inference. *Neuropsychoanalysis* 15, 38–42. doi: 10.1080/15294145.2013.10773716

Friston, K. (2013b). Life as we know it. *J. R. Soc. Interface* 10, 1–12. doi: 10.1098/rsif.2013.0475

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., and Pezzulo, G. (2016a). Active inference and learning. *Neurosci. Biobehav. Rev.* 68(Suppl. C), 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2016b). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

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., Samothrakis, S., and Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biol. Cybernet.* 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. (2017a). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Parr, T., and de Vries, B. (2017b). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

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

Gregor, K., Rezende, D. J., and Wierstra, D. (2016). Variational intrinsic control. *arXiv [Preprint]. arXiv:1611.07507.*

Guckelsberger, C., and Salge, C. (2016). "Does empowerment maximisation allow for enactive artificial agents?" in *Proceedings of the Fifteenth International Conference on the Synthesis and Simulation of Living Systems (Alife 2016)* (Cancun: MIT Press), 8.

Guckelsberger, C., Salge, C., and Colton, S. (2016a). "Intrinsically motivated general companion NPCs via coupled empowerment maximisation," in *Proceedings Conference on Computational Intelligence in Games* (Fira).

Guckelsberger, C., Salge, C., Saunders, R., and Colton, S. (2016b). "Supportive and antagonistic behaviour in distributed computational creativity via coupled empowerment maximisation," in *Proceedings 7th International Conference on Computational Creativity* (Paris).

Guckelsberger, C., Salge, C., and Togelius, J. (2018). "New and surprising ways to be mean: adversarial NPCs with coupled empowerment minimisation," in *Proceedings Conference on Computational Intelligence in Games* (Maastricht).

Hutter, M. (2005). "Universal artificial intelligence: sequential decisions based on algorithmic probability," in *Texts in Theoretical Computer Science. An EATCS Series*, eds W. Bauer, G. Rozenberg, and A. Salomaa (Berlin; Heidelberg: Springer-Verlag).

Karl, M., Soelch, M., Becker-Ehmck, P., Benbouzid, D., van der Smagt, P., and Bayer, J. (2017). Unsupervised real-time control through variational empowerment. *arXiv [Preprint]. arXiv:1710.05101.*

Klyubin, A., Polani, D., and Nehaniv, C. (2005). "Empowerment: a universal agent-centric measure of control," in *The 2005 IEEE Congress on Evolutionary Computation, 2005*, Vol. 1 (Edinburgh), 128–135.

Leike, J. (2016). Nonparametric general reinforcement learning. *arXiv [Preprint]. arXiv:1611.08944.*

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

Little, D. Y.-J., and Sommer, F. T. (2013). Maximal mutual information, not minimal entropy, for escaping the Dark Room. *Behav. Brain Sci.* 36, 220–221. doi: 10.1017/S0140525X12002415

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337. doi: 10.1023/A:1008929526011

Manzotti, R., and Chella, A. (2018). Good old-fashioned artificial consciousness and the intermediate level fallacy. *Front. Robot. AI* 5:39. doi: 10.3389/frobt.2018.00039

Martius, G., Der, R., and Ay, N. (2013). Information driven self-organization of complex robotic behaviors. *PLoS ONE* 8:e63400. doi: 10.1371/journal.pone.0063400

Martius, G., Jahn, L., Hauser, H., and Hafner, V. V. (2014). "Self-exploration of the stumpy robot with predictive information maximization," in *From Animals to Animats 13: 13th International Conference on Simulation of Adaptive Behavior, SAB 2014, Castellón, Spain*, eds A. P. del Pobil, E. Chinellato, E. Martinez-Martin, J. Hallam, E. Cervera, and A. Morales (Springer), 32–42.

Minka, T. P. (2001). "Expectation propagation for approximate Bayesian inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01 (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 362–369.

Mohamed, S., and Rezende, D. J. (2015). "Variational information maximisation for intrinsically motivated reinforcement learning," in *Proceedings Advances in Neural Information Processing Systems 28*, eds C. Cortes, N.D. Lawrence, D.

D. Lee, M. Sugiyama, and R. Garnett (Montréal, BC: Curran Associates, Inc.), 2125–2133.

Orseau, L., Lattimore, T., and Hutter, M. (2013). "Universal knowledge-seeking agents for stochastic environments," in *Algorithmic Learning Theory*, Number 8139 in Lecture Notes in Computer Science, eds S. Jain, R. Munos, F. Stephan, and T. Zeugmann (Berlin; Heidelberg: Springer)158–172.

Ortega, P. A. (2011). Bayesian causal induction. *arXiv [Preprint]. arXiv:1111.0708*.

Ortega, P. A., and Braun, D. A. (2010). A minimum relative entropy principle for learning and acting. *J. Artif. Intell. Res.* 38, 475–511. doi: 10.1613/jair.3062

Ortega, P. A., and Braun, D. A. (2014). Generalized Thompson sampling for sequential decision-making and causal inference. *Complex Adapt. Syst. Model.* 2:2. doi: 10.1186/2194-3206-2-2

Oudeyer, P.-Y., and Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pfeifer, R., Iida, F., and Bongard, J. (2005). New robotics: design principles for intelligent systems. *Artif. Life* 11, 99–120. doi: 10.1162/1064546053279017

Ross, S. and Pineau, J. (2008). "Model-based Bayesian reinforcement learning in large structured domains," in *Proceedings 24th Conference on Uncertainty in Artificial Intelligence* (Helsinki), 476–483.

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

Salge, C., Glackin, C., and Polani, D. (2014). "Empowerment–an introduction," in *Guided Self-Organization: Inception*, ed M. Prokopenko (Berlin; Heidelberg: Springer), 67–114.

Salge, C., Guckelsberger, C., Canaan, R., and Mahlmann, T. (2018). "Accelerating empowerment computation with UCT tree search," in *Proceedings Conference on Computational Intelligence in Games* (Maastricht: IEEE).

Salge, C., and Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Front. Robot. AI* 4:25. doi: 10.3389/frobt.2017.00025

Santucci, V. G., Baldassarre, G., and Mirolli, M. (2013). Which is the best intrinsic motivation signal for learning multiple skills? *Front. Neurorobot.* 7:22. doi: 10.3389/fnbot.2013.00022

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. Auton. Mental Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368

Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. in *Proceedings of the International Conference on Artificial Neural Networks*, Vol. 2 (Perth, WA), 159–164.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA; London: MIT Press.

Toussaint, M. (2009). Probabilistic inference as a model of planned behavior. *Künstliche Intelligenz* 3, 23–29.

Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., et al. (2014). Expectation propagation as a way of life: a framework for Bayesian inference on partitioned data. *arXiv [Preprint]. arXiv:1412.4869*.

Wainwright, M. J., and Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations Trends Mach. Learn.* 1, 1–305. doi: 10.1561/2200000001

Winn, J. and Bishop, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* 6, 661–694.

Conflict of Interest Statement: CG, CS, SS, and DP declare no competing interests. In accordance with Frontiers policy MB declares that he is employed by company Araya Incorporated, Tokyo, Japan.

APPENDIX

A. POSTERIOR FACTOR

Here we want to calculate the posterior factor $\mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi)$ of the complete posterior in Equation (16) without an approximation (i.e., as in direct, non-variational Bayesian inference).

$$\begin{aligned}
&\mathrm{q}(\hat{e}_{\prec t}, \theta | sa_{\prec t}, \xi) \\
&= \frac{1}{\mathrm{q}(s_{\prec t}|a_{\prec t}, \xi)}\, \mathrm{q}(s_{\prec t}, \hat{e}_{\prec t}, \theta | a_{\prec t}, \xi) && \text{(A1)}\\
&= \frac{1}{\mathrm{q}(s_{\prec t}|a_{\prec t}, \xi)}\, \mathrm{q}(s_{\prec t}|\hat{e}_{\prec t}, \theta^1)\, \mathrm{q}(\hat{e}_{\prec t}|a_{\prec t}, \theta^2, \theta^3)\, \mathrm{q}(\theta|\xi) && \text{(A2)}\\
&= \frac{1}{\mathrm{q}(s_{\prec t}|a_{\prec t}, \xi)} \prod_{\tau=0}^{t} \mathrm{q}(s_\tau|\hat{e}_\tau, \theta^1) \prod_{r=1}^{t} \mathrm{q}(\hat{e}_r|a_r, \hat{e}_{r-1}, \theta^2)\, \mathrm{q}(\hat{e}_0|\theta^3) \\
&\quad \prod_{i=1}^{3} \mathrm{q}(\theta^i|\xi^i). && \text{(A3)}
\end{aligned}$$

We see that the numerator is given by the generative model. The denominator can be calulated according to:

$$\begin{aligned}
&\mathrm{q}(s_{\prec t}|a_{\prec t}, \xi) \\
&= \int_{\Delta_\Theta} \mathrm{q}(s_{\prec t}|a_{\prec t}, \theta)\, \mathrm{q}(\theta|\xi)\, \mathrm{d}\theta && \text{(A4)}\\
&= \int_{\Delta_\Theta} \left(\sum_{\hat{e}_{\prec t}} \mathrm{q}(\hat{e}_0|\theta^3) \prod_{\tau=0}^{t} \mathrm{q}(s_\tau|\hat{e}_\tau, \theta^1) \prod_{r=1}^{t} \mathrm{q}(\hat{e}_r|a_r, \hat{e}_{r-1}, \theta^2) \right) \\
&\quad \prod_{i=1}^{3} \mathrm{q}(\theta^i|\xi^i)\, \mathrm{d}\theta && \text{(A5)}\\
&= \sum_{\hat{e}_{\prec t}} \int_{\Delta_\Theta} \mathrm{q}(\hat{e}_0|\theta^3) \prod_{\tau=0}^{t} \mathrm{q}(s_\tau|\hat{e}_\tau, \theta^1) \prod_{r=1}^{t} \mathrm{q}(\hat{e}_r|a_r, \hat{e}_{r-1}, \theta^2) \\
&\quad \prod_{i=1}^{3} \mathrm{q}(\theta^i|\xi^i)\, \mathrm{d}\theta && \text{(A6)}\\
&= \sum_{\hat{e}_{\prec t}} \left(\int \mathrm{q}(\hat{e}_0|\theta^3)\, \mathrm{q}(\theta^3|\xi^3)\, \mathrm{d}\theta^3 \int \prod_{\tau=0}^{t} \mathrm{q}(s_\tau|\hat{e}_\tau, \theta^1)\, \mathrm{q}(\theta^1|\xi^1)\, \mathrm{d}\theta^1 \right.\\
&\quad \left. \times \int \prod_{r=1}^{t} \mathrm{q}(\hat{e}_r|a_r, \hat{e}_{r-1}, \theta^2)\, \mathrm{q}(\theta^2|\xi^2)\, \mathrm{d}\theta^2 \right) && \text{(A7)}
\end{aligned}$$

The three integrals can be solved analytically if $\mathrm{q}(\theta^i|\xi^i)$ are chosen as conjugate priors to $\mathrm{q}(s_\tau|\hat{e}_\tau, \theta^1)$, $\mathrm{q}(\hat{e}_r|a_r, \hat{e}_{r-1}, \theta^2)$, and $\mathrm{q}(\hat{e}_0|\theta^3)$ respectively. However, the sum is over $|\mathcal{E}|^t$ terms and therefore untractable as time increases.

B. APPROXIMATE POSTERIOR PREDICTIVE DISTRIBUTION

Here, we calculate the (variational) approximate predictive posterior distribution of $\mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, sa_{\prec t}, \xi)$ from a given approximate complete posterior. This expression plays a role in multiple intrinsic motivation functions like empowerment maximization, predictive information maximization, and knowledge seeking. For an arbitrary ϕ we have:

$$\begin{aligned}
&\mathrm{r}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \phi) \\
&:= \sum_{\hat{e}_{\prec t}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{r}(\hat{e}_{\prec t}, \theta|\phi)\, \mathrm{d}\theta && \text{(A8)}\\
&= \sum_{\hat{e}_{t-1}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta)\, \mathrm{r}(\hat{e}_{t-1}, \theta|\phi)\, \mathrm{d}\theta && \text{(A9)}\\
&= \sum_{\hat{e}_{t-1}} \left(\int \mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta) \prod_{i=1}^{3} \mathrm{r}(\theta^i|\phi^i)\, \mathrm{d}\theta \right) \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) \\
& && \text{(A10)}\\
&= \sum_{\hat{e}_{t-1}} \left(\sum_{\hat{e}_{t:\hat{T}}} \int \mathrm{q}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, \theta^1)\, \mathrm{r}(\theta^1|\phi^1)\, \mathrm{d}\theta^1 \times \right.\\
&\quad \left. \times \int \mathrm{q}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \theta^2)\, \mathrm{r}(\theta^2|\phi^2)\, \mathrm{d}\theta^2\, \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) \right) && \text{(A11)}\\
&= \sum_{\hat{e}_{t-1}} \left(\sum_{\hat{e}_{t:\hat{T}}} \int \prod_{\tau=t}^{\hat{T}} \mathrm{q}(\hat{s}_\tau|\hat{e}_\tau, \theta^1)\, \mathrm{r}(\theta^1|\phi^1)\, \mathrm{d}\theta^1 \times \right.\\
&\quad \left. \times \int \prod_{\tau=t}^{\hat{T}} \mathrm{q}(\hat{e}_\tau|\hat{a}_\tau, \hat{e}_{r-1}, \theta^2)\, \mathrm{r}(\theta^2|\phi^2)\, \mathrm{d}\theta^2\, \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) \right) && \text{(A12)}\\
&= \sum_{\hat{e}_{t-1}} \sum_{\hat{e}_{t:\hat{T}}} \mathrm{r}(\hat{s}_{t:\hat{T}}|\hat{e}_{t:\hat{T}}, \phi^1)\, \mathrm{r}(\hat{e}_{t:\hat{T}}|\hat{a}_{t:\hat{T}}, \hat{e}_{t-1}, \phi^2)\, \mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}}) \\
& && \text{(A13)}
\end{aligned}$$

From first to second line we usually have to marginalize $\mathrm{q}(\hat{e}_{\prec t}, \theta|sa_{\prec t}, \xi)$ to $\mathrm{q}(\hat{e}_{t-1}, \theta|sa_{\prec t}, \xi)$ with a sum over all $|\mathcal{E}|^{t-1}$ possible environment histories $\hat{e}_{\prec t-1}$. Using the approximate posterior, we can use $\mathrm{r}(\hat{e}_{t-1}|\phi^{E_{t-1}})$ directly without dealing with the intractable sum. From third to fourth line, $\mathrm{r}(\theta^3|\phi^3)$ drops out since it can be integrated out (and its integral is equal to one). Note that during the optimization Equation (47) $\mathrm{r}(\theta^3|\phi^3)$ does play a role so it is not superfluous.From fifth to last line, we perform the integration over the parameters θ^1 and θ^2. These integrals can be calculated analytically if we choose the models $\mathrm{r}(\theta^1|\phi^1)$ and $\mathrm{r}(\theta^2|\phi^2)$ as conjugate priors to $\mathrm{q}(s|e, \theta^1)$ and $\mathrm{q}(e'|a', e, \theta^2)$. Variational inference prediction of the next $n = \hat{T} - t - 1$ sensor values requires the sum and calculation of $|\hat{\mathcal{E}}|^n$ terms for $|\hat{\mathcal{S}}|^n$ possible futures.

C. NOTATION TRANSLATION TABLES

A table to translate between our notation and the one used in Friston et al. (2015). The translation is also valid in many cases for Friston et al. (2016a,b, 2017a). Some of the parameters shown here only show up in the latter publications.

This article	Friston et al. (2015)	Note
$e_t \in \mathcal{E}$		Actual environment states
$\hat{e}_t \in \hat{\mathcal{E}}$	$s_t \in S$	Estimated/modeled environment states
$s_t \in \mathcal{S}$	$o_t \in \Omega$	Actual/observed sensor or outcome values
$\hat{s}_t \in \hat{\mathcal{S}} = \mathcal{S}$	$o_t \in \Omega$	Estimated/modeled (usually future) sensor or outcome values. Note that the index τ instead of t often indicates an estimated future sensor value in Friston et al. (2015).
$a_t \in \mathcal{A}$	$a_t \in A$	Actions
$\hat{a}_t \in \hat{\mathcal{A}} = \mathcal{A}$	$u_t \in U$	Contemplated (usually future) actions
$m_t \in \mathcal{M}$		Agent memory state
$\hat{a}_{t:\hat{T}}$	$\pi, \tilde{u}$	π and $\tilde{u}$ both uniquely specify future action sequences
θ	θ	Generative model parameters
$\mathrm{q}(\hat{s} \mid \hat{e}, \theta^1) = \mathrm{q}(\hat{s} \mid \hat{e})$	$P(o \mid s) = \mathbf{A}_{os}$	Model sensor dynamics, not parameterised in Friston et al. (2015), **A** is a matrix representation
$\mathrm{q}(\hat{e}' \mid \hat{a}', \hat{e}, \theta^2) = \mathrm{q}(\hat{e}' \mid \hat{a}', \hat{e})$	$P(s' \mid s, u) = \mathbf{B}(u)_{s's}$	Model environment dynamics, not parameterised in Friston et al. (2015), **B**(u) is a matrix representation for each possible action u
$\mathrm{q}(\hat{e}_0 \mid \theta^3)$	$P(s_0 \mid m) = \mathbf{D}_{s_0}$	Modeled initial environment state, not parameterised in Friston et al. (2015), **D** is a vector representation. Note, the parameter m is a fixed hyperparameter
$\xi = (\xi^1, \xi^2, \xi^3)$	m	Generative model hyperparam. or model parameter that subsumes all hyperparameters
ξ^1		sensor dynamics hyperparam.
ξ^2		Environment dynamics hyperparam.
ξ^3		Initial environment state hyperparam.
ξ^Γ	(α, β)	Precision hyperparam.
(ϕ, ϕ^Γ)	μ	Variational param.
$\phi^{E_{0:\hat{T}}}$	$\widehat{s}$	Environment states variational param.,
ϕ^{E_τ}	$\widehat{s}_\tau$	for each timestep τ
ϕ^1		Sensor dynamics variational param.
ϕ^2		Environment dynamics variational param.
ϕ^3		Initial environment state variational param.
π	$\widehat{\pi}$	Future action sequence variational param.
ϕ^Γ	$\widehat{\gamma}$	Precision variational param.
$\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$	$\mathbf{Q}(\pi) = \mathbf{Q}(\tilde{u} \mid \pi)$	Variational action-value function. The dependence of $\mathbf{Q}(\tilde{u} \mid \pi)$ on $\widehat{s}_t$ is omitted
$\mathrm{p}(s_{\preceq t}, e_{\preceq t}, a_{\prec t})$	$R(\tilde{o}, \tilde{s}, \tilde{a})$	Our physical environment corresponds to the generative process
$\mathrm{q}(\hat{s}_{\preceq t}, \hat{e}_{\preceq t}, \hat{a}_{t:\hat{T}}, \gamma \mid a_{\prec t}, \xi)$	$P(\tilde{o}, \tilde{s}, \tilde{u}, \gamma \mid \tilde{a}, m)$	The generative model for active inference including γ (which we mostly omit)
$\mathrm{r}(\hat{e}_{0:\hat{T}}, \hat{a}_{t:\hat{T}}, \gamma \mid \pi, \phi, \phi^\Gamma)$	$Q(\tilde{s}, \tilde{u}, \gamma \mid \mu)$	Approximate complete posterior for active inference
$\mathrm{p}^d(\hat{s}_\tau)$	$P(o_\tau \mid m)$	Prior over future outcomes.

Since our treatment is more general than that of Friston et al. (2015) and quite similar (though not identical) to the treatment in Friston et al. (2016a,b, 2017a) we also give the relations to variables in those publications. We hope this will help interested readers to understand the latter publications even if some aspects of those are different. A discussion of those differences is beyond the scope of the present article.

This article	Friston et al. (2016b)	Note
$e_t \in \mathcal{E}$		Actual environment states
$\hat{e}_t \in \hat{\mathcal{E}}$	$s_t \in S$	Estimated/modeled environment states
$s_t \in \mathcal{S}$	$o_t \in \Omega$	Actual/observed sensor or outcome values
$\hat{s}_t \in \hat{\mathcal{S}} = \mathcal{S}$	$o_t \in \Omega$	Estimated/modeled (usually future) sensor or outcome values. Note that the index τ instead of t often indicates an estimated future sensor value in Friston et al. (2015).
$a_t \in \mathcal{A}$	$u_t \in A$	Actions
$\hat{a}_t \in \hat{\mathcal{A}} = \mathcal{A}$	$u_t \in \Upsilon$	Contemplated (usually future) actions
$m_t \in \mathcal{M}$		Agent memory state
$\hat{a}_{0:\hat{T}}$	π,	action sequences
θ	θ	Generative model parameters
θ^1	**A**	Sensor dynamics param.
θ^2	**B**	Environment dynamics param.
θ^3	**D**	Initial environment state param.
ξ	η	Generative model hyperparam. or model parameter that subsumes all hyperparameters
ξ^1	a	sensor dynamics hyperparam.
ξ^2	b	Environment dynamics hyperparam.
ξ^3	d	Initial environment state hyperparam.
ξ^Γ	β	Precision hyperparam.
(ϕ, ϕ^Γ)	$\boldsymbol{\eta}$	Variational param.
$\phi^{E_{0:\hat{T}}}$	$\mathbf{s}_{0:T}$	Environment states variational param.
$\mathrm{q}(\hat{e}_\tau \| \hat{a}_{t:\hat{T}}, a_{0:t-1}, \phi^{E_\tau})$	$(\mathbf{s}^\pi_\tau)_{\hat{e}_\tau}$	For each sequence of actions and for each timestep there is a parameter $\mathbf{s}^\pi_\tau$. Since a categorical distribution is used, the parameter is a vector of probabilities whose entry $\hat{e}_\tau$ is equal to the probability of $\hat{e}_\tau$ if we set $\hat{\mathcal{E}} = \{1, ..., \|\hat{\mathcal{E}}\|\}$
ϕ^1	**a**	Sensor dynamics variational param.
ϕ^2	**b**	Environment dynamics variational param.
ϕ^3	**d**	Initial environment state variational param.
π	$\boldsymbol{\pi}$	Future action sequence variational param.
ϕ^Γ	$\boldsymbol{\beta}$	Precision variational param.
$\hat{Q}(\hat{a}_{t:\hat{T}}, \phi)$	$-\mathbf{G}(\pi)$	Variational action-value function. The dependence of $\mathbf{G}(\pi)$ on $\mathbf{s}^\pi_{0:T}$ is omitted
$\mathrm{p}(s_{\preceq t}, e_{\preceq t}, a_{\prec t})$	$R(\tilde{o}, \tilde{s}, \tilde{a})$	Our physical environment corresponds to the generative process
$\mathrm{q}(\hat{s}_{\preceq t}, \hat{e}_{0:\hat{T}}, \hat{a}_{0:\hat{T}}, \gamma, \theta, \xi)$	$P(\tilde{o}, \tilde{s}, \pi, \gamma, \mathbf{A}, \mathbf{B}, \mathbf{D} \| a, b, d, \beta)$	The generative model for active inference
$\mathrm{r}(\hat{e}_{0:\hat{T}}, \hat{a}_{0:\hat{T}}, \gamma, \theta \| \pi, \phi^\Gamma, \phi)$	$Q(\tilde{s}, \pi, \mathbf{A}, \mathbf{B}, \mathbf{D}, \gamma \| \mathbf{s}^\pi_{0:\hat{T}}, \boldsymbol{\pi}, \mathbf{a}, \mathbf{b}, \mathbf{d}, \boldsymbol{\beta})$	Approximate complete posterior for active inference
$\mathrm{p}^d(\hat{s}_\tau)$	$P(o_\tau) = \sigma(\mathbf{U}_\tau)$	Prior over future outcomes.

HYPOTHESIS AND THEORY
published: 26 October 2018
doi: 10.3389/frobt.2018.00121

Can Computers Become Conscious and Overcome Humans?

Camilo Miguel Signorelli[1,2,3]*

[1] Department of Computer Science, University of Oxford, Oxford, United Kingdom, [2] Cognitive Neuroimaging Unit, INSERM U992, NeuroSpin, Gif-sur-Yvette, France, [3] Centre for Brain and Cognition, Pompeu Fabra University, Barcelona, Spain

The idea of machines overcoming humans can be intrinsically related to conscious machines. Surpassing humans would mean replicating, reaching and exceeding key distinctive properties of human beings, for example, high-level cognition associated with conscious perception. However, can computers be compared with humans? Can computers become conscious? Can computers outstrip human capabilities? These are paradoxical and controversial questions, particularly because there are many hidden assumptions and misconceptions about the understanding of the brain. In this sense, it is necessary to first explore these assumptions and then suggest how the specific information processing of brains would be replicated by machines. Therefore, this article will discuss a subset of human capabilities and the connection with conscious behavior, secondly, a prototype theory of consciousness will be explored and machines will be classified according to this framework. Finally, this analysis will show the paradoxical conclusion that trying to achieve conscious machines to beat humans implies that computers will never completely exceed human capabilities, or if the computer were to do it, the machine should not be considered a computer anymore.

Keywords: artificial intelligence, information processing, cognitive computing, type of cognition, super machine, conscious machine, consciousness

OPEN ACCESS

Edited by:
Angelo Cangelosi,
Plymouth University, United Kingdom

Reviewed by:
Rodolphe Gelin,
Softbank Robotics, France
Junpei Zhong,
National Institute of Advanced Industrial Science and Technology (AIST), Japan

****Correspondence:***
Camilo Miguel Signorelli
camiguel@uc.cl

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *26 August 2017*
Accepted: *26 September 2018*
Published: *26 October 2018*

Citation:
Signorelli CM (2018) Can Computers Become Conscious and Overcome Humans? Front. Robot. AI 5:121. doi: 10.3389/frobt.2018.00121

INTRODUCTION

During many centuries, scientists and philosophers have been debating about the nature of the brain and its relation with the mind, based on the premise of an intrinsic dualism, typically called mind-body problem (Searle, 1990; Chalmers, 1995). Arguments take one form or another, however, most of them can be reduced to one kind of dualist or non-dualist view (Lycan and Dennett, 1993). The importance of these debates acquires even more relevance when the question is stated as the possibility to build machines which would be able to reproduce some human capabilities such as emotion, subjective experiences, or even consciousness.

The problem is exacerbated when some scientists claim a new future generation of computers, machines and/or robots which would additionally overcome human capabilities. In the view of the author, these claims are based on misconceptions and reductionism of current most important issues. The idea, however, is not discarded here and is expressed, trying to avoid reductionism, in a different way to show its paradoxical consequences (Signorelli, 2018). For example, the idea of reaching and overtaking human capabilities implies the knowledge of a set of distinctive processes and characteristics which define being a human (e.g., intelligence, language, abstract thinking, the creation of art and music, emotions and physical abilities, among others). This simple idea leads to some fundamental issues. First, claims about new futurist robots do not define this set of distinctions; they do not care about the importance of what it is to be a human, what is necessary to build conscious machines or its implications. Secondly, they assume a materialist view of these distinctions (i.e., these distinctions emerge from the physical and reproducible interaction of matter) without explaining the most fundamental questions about the matter (Frank, 2017). Thirdly, they do not explain how subjective experience or emotions could emerge from the theory of computation that they assume as a framework to build machines, which will reach consciousness and overcome humans. In other words, these views do not explain foundations of computation that support or reject the idea of high-level cognitive computers. Finally, engineering challenges of building these kinds of machines are not trivial, and futurists assume reverse engineering as the best tool to deal with this when even some neuroscience techniques do not seem to give us any information about simple computing devices such as microprocessors (Jonas and Kording, 2017). Actually, if methods of neuroscience are not inferring useful information from microprocessors, it is possible to conclude that either the neurons are not working as computers or all the information that we know about cells and neurons, using these techniques, is wrong. The first option discards reverse engineering as a feasible tool to understand the brain, and the second option discards findings in neuroscience related to mechanistic and computational interpretation. Thus, it is still necessary to focus on many intermediate and fundamental steps before declaring that some computers would reach or even exceed human capabilities.

This work does not expect to solve these issues; on the contrary, the aim of this paper is to expand previous works (Signorelli, 2018) and illustrate misconceptions and misunderstanding of some crucial concepts. For example, the issue of overcoming human capabilities will be discussed in parallel with the issue of producing conscious machines, to show their close relation and same paradoxical consequences. Additionally, the importance of new concepts and ideas will be approached in a preliminary and speculative way, with the intention of developing them in further works. Following this framework in order to make clear some of the questions above, the second section will define what will be understood by human capabilities and human intelligence; the third section will confront current common views of computation, cognitive computing, and information processing; the fourth section will discuss consciousness as a basic requirement to make computers with similar human intelligence; the next two section will show a new hypothesis of how consciousness could work; then, machines will be classified in four categories based on four types of cognitions derived from consciousness requirement, and finally, according to these classifications, the last section will show some paradoxes and implications, which emerge from the idea to make machines-like-brains reaching consciousness and overcoming humans.

A SUB SET OF HUMAN CAPABILITIES

Usually, it is considered that computers, machines and/or robots will eventually reach, or even overtake human intelligence. This idea is supported by many advances in Artificial Intelligence (AI). For example, consecutive victories of DeepMind project vs. the GO human champion (Silver et al., 2016), or robots that have passed some kind of Self-Consciousness test (Bringsjord et al., 2015). Science fiction, movies, and writers also stimulate and play enough with the notion of "Singularity," the precise moment where machines exceed human capabilities (Good, 1965). In this scenario, a computer/machine is called Super Machine.

Nevertheless, how much does scientific evidence support this idea? What does overcoming human intelligence mean? What does human intelligence mean? And what is the relation with consciousness? Computers already exceed human algorithmic calculations, among many others. A clear example is the recent report of AlphaGo zero which can learn without human intervention and play at super-human level (Silver et al., 2017). In fact, one option to overcome human abilities might be a cognitive system completely different to the anthropocentric science fiction view. As will be shown later, this kind of computer may reach and overcome some, but not all, human capabilities. That is why; one position could claim that it is not necessary to assume computers like brains or conscious machines to overtake human capabilities. It is a valid point; however, will this kind of computer surpass human brain only in a rational/algorithmic way or also an emotional one? Will this kind of computer be able to dance better than us, to create better than us, to feel better and like us? Otherwise, it will never reach nor overtake human abilities. One reason is that part of being human is to have emotional behavior, to be able to dance, create, etc, additionally to our apparently rational behavior. As it was mentioned above, the first issue emerges: what does human being mean? If what is being a human and which abilities need to be overcome are not understood, how can we ever think about overcoming unknown capabilities? For example, human intelligence may not be only associated with logical, algorithmic, or rational thinking. Types of intelligence have already been suggested, which are closely related to each other such as kinaesthetic and emotional intelligence in humans (Sternberg, 1997; Gardner, 1999). So far, implementing emotions or simple movements in machines is equal to or more complicated than implementing rational or algorithmic intelligence (Moravec, 1988). Actually, current implementations of emotions in machines are based on a logical, computable and deterministic approaches, leaving out essential characteristics of emotions such as that emotions interfere with rational processes and optimal decisions. In fact, these implementations are founded on the idea that emotions play an important role in making humans more efficient, rationally speaking (Martinez-Miranda and Aldea, 2005), when cognitive fallacies are showing the contrary (Gilovich et al., 2002; Kahneman, 2003) and experiments on neuroscience from the called default neural network, which is related to self-oriented information, are suggesting anti-correlated subsystems of information processing (Simpson et al., 2001; Fox et al., 2005; Buckner et al., 2008) which interfere each other. The view of computer non-like-brain does not care about these issues and assumes intelligence as only rational, logic and computable capability; or even worst, the problem of computer non-like-brains defenders is to think that some properties of life could be replicated without the distinctive properties of being alive.

Then, is it possible to define a set of human characteristics? Futurists assume the existence of this set but they do not define it in any way. While any serious attempt to define a human set or a subset should first start with a definition of living entities. One possible definition is the notion of autopoiesis (Maturana and Varela, 1998) which refers to the self-reproduction and self-maintenance of a system. In this view, a living machine is a unitary system or network of processes which is able to regenerate through their interactions and continuous transformation. Even when it is still controversial a complete definition of living beings and the utility of the autopoiesis concept (Fleischaker, 1992), two characteristics, autonomy and reproduction, emerge as key features of living beings. Some critics of this concept state that autopoiesis does not consider external references that can be crucial for the organism. Therefore, a probable better definition of the living being may be a unitary system or network of processes which interacts with the environment to keep their autonomy and increase their capability to reproduce. Of course, any definition of life is a huge enterprise and the goal of this essay is not to answer this question, but state a simple and probably the simplest definition that can help us to decide when a machine reaches and overcomes human characteristics. Interestingly, this general definition does not discard the idea that other systems or machines can reach these two characteristics, even when they should not be considered living machines. In fact, this is not contradictory because autonomy and reproduction are thought here as a subset of living machine properties; it means that they are necessary conditions but not sufficient to be considered living machines. Thus, humans, as well as other animals, are autonomous entities with the ability to reproduce.

Additionally, however, it is also necessary to identify at least one characteristic to differentiate human being from other living beings. One historical proposal has been the notion of morality. Morality and ethics can be understood as high-level reasoning to distinguish between proper or improper behavior and intentions. This notion also implies a community, a culture and social obligations within that community. Morality has been studied by many philosophers as for example (Hegel, 2001) and (Kant, 1785), and connected with concepts as rationality, free will, and consciousness. Nevertheless, when neuroscientists look for correlates or building blocks of morality inside of the brain, it is possible to find areas which are associated with empathy and social interaction, mostly identified with emotional states (Bzdok et al., 2012). In these terms, morality is not only a rational process as some philosophers proposed (Kant, 1785), and it is apparently not exclusive of human beings. Thus, the notion of a uniquely human characteristic remains too elusive and what it is necessary to explain and replicate in robots is still not clear (Chappell and Sloman, 2007). That is why; the suggestion in this work is to define human morality as a complex process where rational and emotional thinking takes part, then, moral decisions, moral behavior, and moral intentions emerge only after this intricate process takes place. In other words, the distinctive ingredient in human intelligence will be considered the capability to integrate rational and emotional thinking to take moral decisions which are adapted to the context. It is not clear that animals can integrate, as a whole, rational thinking and emotional thinking to take moral decisions, however, even if some animals could be able to do it, the assumption here is that the kind of morality emerged would be different and characteristic of each species, culture and even subjects. In other words, as it will be shown later, morality is a complex behavior intrinsically related to context, subjectivity, and consciousness.

The definition of a general intelligence can also be inferred from the previous discussion, at least in a preliminary way. It is interesting to point out that a general definition of intelligence and human intelligence is still a question of debate, since the pioneering works of Turing (Turing, 1950) until our days, where the definition changes according to how science and AI evolve (Nilsson, 2009; Stone et al., 2016). Nevertheless, based on previous comments, general intelligence can be understood as the capability of any system to take advantage of their environment to achieve a goal. Biologically speaking this goal is maintaining the autonomy and reproduction, that is to say: survive; while the goal in machines can be solving a specific task or problem using internal and external resources. This general definition can incorporate living beings as well as robots and computers, and in this way, intelligence is general enough to include different kinds of intelligence, contextual influences and different kind of systems with different degrees of intelligence. Finally, also in these terms, human intelligence would be the ability to take advantage of their social environment to keep autonomy and reproduction thanks to a balance between rational and emotional information processing. This human intelligence definition incorporates the set of distinctive characteristics which define partially being human, and where the advantage can take place through cognition, learning, memory; among other processes needed to achieve the goal.

At this point is inevitable to shortly mention something about potential tests to prove if a machine reached or not the criteria of human intelligence. Turing was the first one to suggest a test based on a simple exchange of words, questions and answers (Turing, 1950). In its simplest version, this exchange is between a machine and a human who should decide if the machine is a machine or another human. The test is simple, in the sense of its simple execution, and at the same time complex, in the sense that it should capture as many as possible features of the human being. Turing probably realized that the complexity of human intelligence was not only associated with rational and logical processes. That remains evident in the way as he proposed his test as a simple written conversation and also when he refers to the incorporation of human mistakes in future machines to be able to pass the test. However, the Turing test has been criticized many times, where the main against argument is summarized by Searle in his Chinese room example (Searle, 1980). A full review of this topic would be part of an entirely new document and indeed, it will be part of further works. In this way and from the definition of human intelligence stated above, it seems better to suggest a test founded on moral dilemmas more than simple day to day questions (Signorelli and Arsiwalla, 2018). Moral dilemmas are simple, in the sense that they do not require any kind of specific knowledge, but at the same time very complex even for humans, because some of them require a deep understanding of each situation, and deep reflexion to balance moral consequences, emotions, and optimal solutions. No answer is completely correct, they are context dependent, and solutions can vary among cultures, subjects, or even across the same subject in particular emotional circumstances. In other words, a moral test, grounded on moral thinking, needs intermediate processes which are characteristics of high-level cognition in human, as for example self-reflection, sense of confidence and empathy, among others. Hence, a machine will reach part of what it is defined as human intelligence if the machine is able to show autonomously speaking the intricate type of thinking that humans have when they are confronted to these kinds of dilemmas. To do that, it is necessary to focus on intermediate steps reaching some of the previous processes of moral thinking in humans (**Figure 1**).

One example of a moral test is the next situation (**Figure 1A**): If you are in an "emergency boat" after a shipwreck and the boat has only one space left, who would you admit to in the boat and why: a big, healthy and young dog or an injured and sick old man? The answer is not obvious and actually, it is one of the most debated topics in biomedical research, because it does not involve only human morality but also inter-species issues on animal experimentation. What could be the answer of a machine to this question? What could be the logical and emotional thinking of this machine? What is, in fact, the answer of the reader? There are very good reasons to take any of both possible decisions, even a third and fourth answer is also possible, however, the important point is the way how to reach to a conclusion and not the conclusion itself. Of course, many critics should be addressed before to claim that a moral test would be a good test to capture the machine intelligence, compared with human intelligence. For example, according to what types of answers will the comparison be made? What would happen if the machine develops its own sense of morality? Will we be able to recognize it? Tests for machines apparently make sense only when it is desirable to compare them with human intelligence, but in fact, if the machine reaches consciousness, it is also possible that the machine develops a new kind of morality based on non-anthropocentric views and even new possible answers to many moral dilemmas.

For the purpose of this work, we will need to assume that there is a certain set of "human being" properties formed by at least a subset of three features: Autonomy, Reproduction, and Morality. Therefore, it is possible to decide when an animal or machine reach or not the condition to be part of this set, even though it is known that the definition of this set is one of the most controversial and debated issues. Moreover, to reach these three main elements it is necessary to incorporate many intermediate steps and some of them will be discussed in next sections. For example, robots and computers are rarely autonomous in the biological sense; they definitely cannot replicate, re-structure or even recover from harm by themselves. However, these issues can be overcome in the future, at least in a functional way. The only huge issue that is not possible to implement without a deeper understanding of human beings is the morality question, paradoxically, an important distinctive human characteristic, closely to human intelligence and consciousness. Morality requires many previous processes usually considered as high-level cognition, starting with decision-making to self-reflection, to be able to detect mistakes on these decisions; sense of confidence, to estimate how correct a decision or action is; mental imagery, to create new probable scenarios of action; empathy, to equilibrate individual and social requirements; understanding of context, to adapt moral decisions to the context, among others. Because these processes are sharply connected with consciousness, as it will be shown in next sections, a moral test is also a kind of consciousness test. Until now, brains are the only types of systems that have these processes and focusing on how they are working will help us to understand what it would be necessary to replicate in robots for them to reach consciousness and potentially achieve high-level cognition.

Further work and potential experiments can be influenced by these preliminary ideas, in order to improve the behavior of robots/machines trying to answer what is necessary to replicate a truly moral behavior in them.

INFORMATION PROCESSING IN THE BRAIN

One supporting fact about the idea of reaching consciousness and overcoming human capabilities with computers comes from the exponential increase of computational capacity or Moore's law (Moore, 1998). This increase should impact on the development of new technologies until reaching intelligence levels of the human brain. Beyond this view, there is the assumption that the brain works as a computer and its processing could work by analogy with computational processes. Of course, the brain is a physical entity as computers are; it partially works with electrical signals, resolves complex problems and is processing information in one way or another. Nevertheless, the way the brain processes information is still unknown and, it may not be a digital computation, or rather not be information processing in computational abstract terms at all (Epstein, 2016). Information processing implies processes where input are changed to become outputs; however the brain could be working in a new regime, where the distinction between inputs and outputs could not exist, even causalities could be completely different to what we know until now. In this context, it should be possible to speak about another kind of processing as "replication processing," "simulations" (Arsiwalla et al., 2018) or maybe "abstract models," which could be self-informative to some singular physical systems like brains. It is also known that brains work with complex neuromodulation (Nusbaum et al., 2001), stores information in a sparse and unknown way (Tetzlaff et al., 2012; Gallistel and Balsam, 2014), and most distinctive yet: complex properties as subjective experiences, emotions, consciousness (Cleeremans, 2011; Dehaene et al., 2014; Tononi et al., 2016) and biased behavior (Ellsberg, 1961; Gilovich et al., 2002; Moore, 2002; Machina, 2009) emerge from the brain. These emergent properties do not have any obvious correlation with higher or lower computational capability. For example, the cerebellum has more neurons than any other part of the brain,

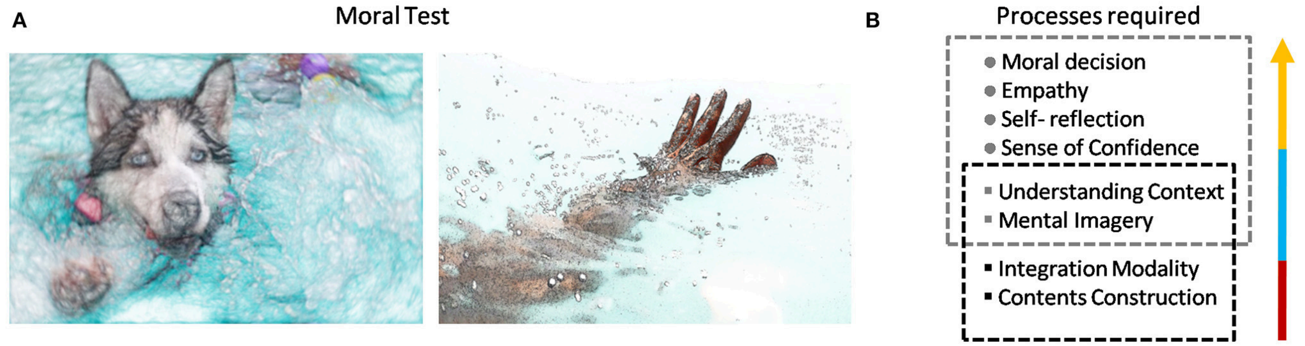


FIGURE 1 | Moral Test and Processes required. **(A)** Moral test and moral dilemmas are suggested to test when a machine has reached human kind of thinking. **(B)** Some processes required for moral thought are stated as examples, among many other possible processes needed.

but it does not play any important role in conscious perception (Tononi and Koch, 2015).

Related to this notion, a common assumption in cognitive science is to consider the processing of information as a synonym of computation; however, it is necessary to differentiate both concepts. For instance, if the information is considered as the content of a message, this content would need a physical system to be propagated and stored. Thus, information may be understood or at least associated with a physical entity (Landauer, 1999). According to a general view, information processing can be any physical process which transforms an input into an output. Information processing can also be defined in terms of causality between inputs and outputs. Additionally, computation is mainly understood as syntactic and symbolic manipulation of information (Searle, 1990). In this sense, computation is an algorithmic and deterministic type of information processing. Although it is possible to appeal to a non-deterministic computation, in general, this non-deterministic computation can be reduced to deterministic types of simple computation at the level of a Turing machine. The problem is that brains are not just doing computation, they are also able to give interpretations and meaning to their own high-level information processing. Arguments in favor of this idea are stated from philosophical view in Searle (1990) and psychological/biological view in Cleeremans (2011).

One interesting case of computation is artificial neural networks, which could be interpreted as semi-deterministic information processing systems. Artificial neural networks evolve in a non-deterministic way thanks to self-learning and training from some given rules, which are not always explicitly programmed. These systems are semi-deterministic in the sense that it is not always possible to ensure what the net is learning, nor control the dynamic evolution of its learning process, even if deterministic learning rules have been given. Of course, it is in part because of the noise or randomness of the training data set, and/or due to predominant statistical features of the data set that were not well controlled. However, even if all these properties are controlled, it is never known what the network has learned until it is tested and even after testing; it is never possible to be sure about which node or layer encodes one or another statistical property of the data. Actually, it looks more like a domain-global and distributed characteristic than local (Christian et al., 2014). Therefore, it is not possible to fully determine or predict classically speaking the way how the net will behave. Neural and artificial neural nets are neither completely indeterminate nor determinate, but semi-determinate. Since artificial neural networks, as for example Hopfield networks (Hopfield, 1982), are inspired by biological principles (Hebb, 1949; Gerstner et al., 2012), which are in turn inspired by biological observations (Caporale and Dan, 2008), one option to introduce the semantic and meaning to artificial networks would be the implementation of interactions between subsystems as observers of each other in a context of artificial neural networks. This will be discussed in section five. Through this way, intelligence would not be only associated with deterministic logical computation but with the interaction between deterministic, semi-deterministic, non-deterministic, and perhaps quantum computation/simulations, or even new frameworks of processing of information.

While some computer and cognitive scientists might not agree with this interpretation of information and computation, it is still admissible to have processing of information without computation and intelligence without a deterministic way of processing of information. Actually, the brain apparently does it. In fact, the most important features of the brain are the result of unpredictable, nonlinear interactions among billions of cells (Ronald and Nicolelis, 2015; Haladjian and Montemayor, 2016). Science does not know the real "language" of the brain; does not know how cognitive abilities emerge from physical brains, and even more complicated, it is not certain that we have a deterministic way to explain how this emergence works.

At this point, the usual idea of digital computation in cognitive science and neuroscience should change in favor of a perspective of computation and information processing by analogy with physical systems where inputs, rules and outputs can be interpreted in a physical and global way.

The brain should not be thought as a digital computer neither in the "software" (Searle, 1990; Chalmers, 1995) nor in the "hardware" (Llinas et al., 1998; Bullock et al., 2005; Epstein, 2016). One reason is that this analogy obscures the complex physical properties of the brain. On the one hand, neuroscience and cognitive science use indiscriminately concepts as information, computation and processing of information without understanding their physical counterpart, sometimes based on the assumption of non-hardware dependency of these concepts, other times because of the assumption that the brain encodes and decodes information (and how it does so). The most common assumption is to think that activation or spikes in neurons are the only informative state. While other cells, for example astrocytes (Alvarez-maubecin et al., 2000), and non-classical integration such as neuromodulatory substances (Nusbaum et al., 2001), back-propagation (Stuart et al., 1993), among others (Bullock et al., 2005) are ignored. In addition, inactivation and deactivation states could also carry valuable information about dynamical brain states at macro and micro scale. Neurons are never in a static state and their membranes are presenting fluctuations that could still be informative (for instance, Sub-threshold oscillations). The distinctive physical brain properties and their dynamical interactions are apparently more important than in digital interpretations, what implies that hardware cannot be ignored at all. According to this point, the analogy between a drum and the brain would be more relevant than the analogy brain-computer. Drums can respond with different and complex vibration states when they are stimulated, and they can be also understood on computational terms: input (hits), rules (physical laws, physical constraints such as material, tension, etc.), and outputs (vibration, sounds, normal modes). Indeed, the brain has many more similarities with a dynamical system as a drum than with digital computers, which are based on discrete states. Drums, as well as brains, are dynamical systems with emergent and sub-emergent properties, drums have different modes of vibration, superposition, physical memory, sparse "storage" of this memory, among others features. In abstract terms, drums are also "computing" and processing information, but this information processing is a dynamical reaction from external/internal stimuli more than a formal calculation process (computation as defined above).

On another hand, computer science is missing valuable information on the attempt of replicating brain capabilities. One example is alpha, gamma or oscillations of brains in general (Buzsáki and Draguhn, 2004), synchrony (Varela et al., 2001; Uhlhaas et al., 2010), harmonic waves (Atasoy et al., 2016), among other processes which are not seriously considered in artificial intelligence, not even using artificial neural networks. Sub-emergent properties in the brain may be also important, such as plasticity changes due to the intentional practice of meditation (Lutz et al., 2004; Brefczynski-Lewis et al., 2007). These characteristics should be understood and incorporated in order to implement the social behavior in new generations of computers, machines and robots. Considering that some of these behaviors are intrinsic to biological organisms, perhaps these behaviors are not reproducible without some intrinsic constituents of information processing of biological organisms (Chappell and Sloman, 2007; Sloman, 2007) as for example oscillations or neurotransmitters.

Finally, abstractions and general concepts are really useful in theoretical terms; however, concepts as computation, information, and information processing in the brain do not have evident interpretation. Realizing that these concepts should not be used as an analogy with computers is the only way to lead us to the correct direction: Focusing on differences between brains and computers, and trying to fill the gaps without assumptions. Maybe, for many computer scientists, these comments are trivial, but what computation means for computer science is not the same as for biological science, leading to misunderstandings and misconceptions, while also the knowledge that computer sciences have about "codification" in the brain is very limited, leading to erroneous assumptions.

To sum up, sections two and three have identified some usual presumptions: (i) The assumption of a set of distinctive properties defining human being without focus on the distinctive properties of human being, (ii) intelligence related only to logical and rational thinking, (iii) brains working by analogy with hardware-independent computers, (iv) computation as synonym of information processing, and (v) brain information only "encoded" in the activation states of neurons. When differences between concepts appear, it becomes necessary to clarify some of them. That is why a subset of the features of human beings has been identified and some concepts clarified. For example, a better understanding, and definition of information processing in the context of human intelligence, where computation will be a kind of information processing among many other types, including the characteristic one to biological organisms (Chappell and Sloman, 2007). Probably, new concepts and foundations of information will be also needed, especially to understand the real language of brain cells, as a crucial theoretical starting point. These foundations should be inherent to minimal constitutive parts of physical theories and as it mentioned above, important hardware requirements, emergent, plasticity and sub-emergent properties should be considered in any attempt to replicate brains features. Thus, a computer-brain metaphor is not useful anymore, at least in the current sense. Nevertheless, it could still be possible to replicate some brains abilities thanks to new formulations of information processing and theoretical frameworks.

CONSCIOUSNESS AS REQUIREMENT FOR HUMAN INTELLIGENCE

Intelligence should also be considered as a whole. Intelligence is often understood as the ability to solve problems in an efficient way, thanks to other mechanisms like learning and memory. It means the maximization of the positive results in a certain solution while minimizing the negative impacts, for instance, waste of time. To do that, other processes, such as learning and memory, are also needed and associated with the definition of intelligence. In a general sense, learning has been understood as the process to gain new knowledge or improve some behavior, while the memory is the storage of this knowledge. To solve problems efficiently, it is necessary to access a certain memory that was acquired thanks to a specific learning that will modify again the memory of the system. The more intelligent is the system, the more it learns. However, in that framework, it is forgotten that emotions, subjective experiences, and cognition are deeply connected with human intelligence (Haladjian and Montemayor, 2016). They play a crucial role in learning, in the consolidation of memories, in retrieved memory and human cognition in general (Cleeremans, 2011).

Therefore, as it was stated in section A Sub Set of Human Capabilities, intelligence is better defined as the capability of any system to take advantage of their environment to achieve a goal. Specifically, human intelligence would be the ability to take advantage of their environment to keep autonomy and reproduction thanks to a balance between rational and emotional information processing. With this last definition, both main features on human thinking, reason and emotion, are merged in one global concept, together with two other features, autonomy and reproduction, that also define, altogether, the potential set of human being properties. In this context, perception, cognition, learning, and memory are key features of human intelligence considered as a whole and emerged from specific soft properties of brains, such as for example neural plasticity and oscillations. Learning and memory are intrinsically dynamic processes in the brain, changing all the time and conditional to these soft neural properties, while for computers, memory is a very static feature, mainly grounded on symbolic discretization, and in the best case, learning is driven for efficient algorithms which are also statics. Biologically, the more intelligent the system, the more balance the system has between different inner processes to achieve specific or general goals. For example, a computer is designed to make faster calculus, algorithms, and other kinds of very useful tasks, however, the computer cannot take advantage of anything that it does, in conclusion, computers are not really intelligent. Nevertheless, the last version of AlphaGo zero (Silver et al., 2017) can learn by itself and take advantage from the knowledge given as input, to improve its own performance in a specific task, as for example playing Go. Using the intelligence definition stated here, this system is more intelligent than a simple computer. By analogy, if a lizard is compared with a mouse, the later has a larger repertoire of actions, taking more advantage of their environment, than the lizard. In this sense, mice are more intelligent than lizards. It is possible to continue and even define which humans will be "more intelligent" than others looking at how they take advantage of the environment in a way that they balance both rational and emotional costs. For instance, a person who wins a discussion with his partner at the expense of their relationship is less intelligent than who wins the discussion and keep a good relationship. The crucial point is that emotions are playing an important role in classical processes of natural intelligence such as learning and memory, but they are also playing a crucial role increasing the repertoire of actions and possibilities to achieve biological goals. These new behaviors are not, paradoxically, always efficient, in a logical way, but they are the best way to achieve the goal according to the system strategy (learned by experience) even when they can interfere with rational/optimal solutions. Emotions are not just used to improve memory or learning curves; they are also useful to increase the variability and unpredictability of behavior.

Furthermore, one requirement for emotional and logical/rational intelligence, as starting point to show some of the subset human features mentioned above, seems to be what is called subjective experience (Barron and Klein, 2016) or in a more complex order: Consciousness. On the one hand, high level processes needed for moral thinking such as self-reflection, sense of confidence, error detection, understanding context, among others (**Figure 1B**) are essential part of consciousness and subjective experience as a whole (Gehring et al., 1993; Smith, 2009; Fleming et al., 2012). Self- reflection and sense of confidence are understood as the ability to report a mistake, like error detection, and grade the confidence of some decisions or action, even before receiving any feedback about the mistake. In fact, some researchers have suggested the intrinsic relation between social complexity associated with these processes and the emergence of consciousness (Arsiwalla et al., 2017). On another hand, humans first need to be conscious to take some complex rational decisions, to plan, and to have the intention to do something (Baars, 2005; Tononi and Koch, 2008). For example, vegetative patients and minimally conscious patients do not present signals neither planning nor having intentions to do minimal tasks (Gosseries et al., 2014), even when they could present minimal signs of consciousness (Owen et al., 2006). Planning and intentions apparently emerge when minimal signs of consciousness exceed a threshold. In fact, these minimal signs can be interpreted as predictors of recovering in minimally conscious patients (Bekinschtein et al., 2009; Casali et al., 2013). Other works are re-defining the idea of subjective experience until its minimal constitutive part and argue the existence of basic subjective experience even in insects (Barron and Klein, 2016). It would mean that complex decisions, planning, and have intentions which are needed to moral thoughts are different from consciousness, although they are closely related: Subjective and conscious perceptions are apparently previous to rational intelligence, planning, moral thoughts, and even efficient behaviors. For example, experiments in the psychology of judgment and behavioral economics have also shown that subjects tend to perform some tasks in a biased manner even if they have been trained, suggesting that logical and rational intelligence appear only after more elaborated information processing (Gilovich et al., 2002; Kahneman, 2003). It is clear that how biology implements high-level intelligence is completely different from how computer science implements it (Moravec, 1988). The whole set of human intelligence, as the capacity to take advantage of the environment, would only emerge after awareness.

The need to incorporate subjective experience and eventually consciousness to reach complex intelligence implies a complex problem which involves many different processes as awareness, emotions, subjectivity, intentionality, and attention, among others. Consciousness should be composed by all of these processes like a differentiated and unified whole, but it is not any of them. For example, it could be necessary to be aware to have emotions and subjective experiences, or maybe vice versa, and we will need them to show intentionality, attention and high-level cognitive abilities. It is also necessary to insist and distinguish that these are different processes, for instance, awareness and attention; while it is important understanding all of them as constituent parts of what we describe as consciousness. For example, at least two main processes have been identified with consciousness: (1) the fact of knowing something or what here will be understood as awareness, i.e., to become aware of something and/or perceive something internally or externally, and (2) to know that I know or do not know something, or more precisely the notion of self-conscious systems (Varela, 1975) as a "monitoring" process of this awareness and connected with the more general concept of self-reference (Varela, 1975; Kauffman and Varela, 1980; Kauffman, 1987). It is worth differentiating self-reference, as an autonomous process (where a third system emerge from its own interactions; Goguen and Varela, 1979), from other interpretations, as for instance self-monitoring as control process (where a second or third system, at the same "complex" level than others, is needed to control; Dehaene et al., 2017). Here, the notion refers to the idea of self-reference for living machines. Thus, awareness is also understood as conscious or non-conscious "contents" and self-reference is connected with conscious or non-conscious manipulations (processing) of "contents" (Shea and Frith, 2016), or what will be more precisely called "neural objects." In this sense, subjectivity and conscious perception apparently needed to reach rational, emotional, and moral thoughts are associated with awareness and self-reference as crucial ingredients of consciousness. Nevertheless, consciousness is not reduced to the possible relationship between awareness and self-reference, it is the whole process of processes interconnected with awareness, self-reference, subjectivity, rational and emotional thoughts, among many others. Consciousness emerges from all of them as a whole (Varela and Goguen, 1978). Hence, after consciousness emerges from the interaction between these processes, human intelligence would appear as the group of strategies to take advantage of the environment thanks to the balance of emotional and rational information processing.

Four types of cognition and some of their associated tasks can also be defined from awareness and self-reference (Shea and Frith, 2016; Signorelli, 2017; **Figure 2A**): (1) Type 0 Cognition corresponds to systems which have neither awareness of their internal or external contents nor self-reference of their internal processes. One example in humans is motor control. Motor control is the automatic control that the neural central system has to move some joints and muscles without any necessity of voluntary control or awareness. Many apparently high-level tasks in human can be classified in this category, as for example the extraction of individual word meaning and primary attention sometimes called priming. (2) Type 1 Cognition is defined as the type of cognition emerged when a system is aware of their contents. In other words, it is aware of the elements that the system needs to manipulate and solve particular or general problems, but the system does not monitor this manipulation. It can be also associated with a holistic kind of information. For example, when subjects answer very quickly to some apparently intuitive questions but their answers are normally wrong (Fallacy questions). Type 1 cognition also involves mental imagery, emotions, voluntary attention and most of our subjective capabilities as to be aware of the experience of color or pain, among others. (3) Type 2 Cognition appears when the system is aware of their contents and also has self-reference capability as the ability to manipulate them. This type of cognition involves the high-level cognitive capabilities defined above and needed for human morality. Some tasks, which are part of this type of cognition, can be: the ability of self-reflection; rational thinking; detection of error even before receive any clue about the mistake; sense of confidence, before and after any decision; complex meanings; voluntary and quick learning, among other interesting features of human thinking. (4) Finally, Type ∞ cognition incorporates the manipulation of contents without awareness of their contents. In other words, the system has self-reference, but it cannot extract meaning either from their manipulation nor their contents. It could be like an automaton, and actually, there is not a biological example of this category.

These categories will help us to classify the kind of machine and the characteristics needed as a requirement to reach or overcome human cognitive capabilities. These ideas may imply that to reproduce high-level of human intelligence following biological principles, it is necessary but not sufficient to introduce first, subjective and conscious behavior in machines at early stages to reach the type 1 and type 2 cognition of human beings. Then, the question of overcoming humans is intrinsically related to the question of build conscious machines. In this way, machines will be classified by analogy to the cognitive level that can reach according to the types of cognition emerged from awareness and self-reference (**Figure 2B**). These two processes would be previous to complex kind of cognition, as for example type 2 cognition, voluntary learning and complex memories, but only sufficient features to overcome humans if autonomy, reproduction, and morality are also reached. In other words, the only way to reach human brains would be making conscious machines capable of reproducing emotional human intelligence, in addition to logical intelligence, and keeping their autonomy, reproduction capacity, and reaching moral/ethical thinking. Otherwise, machines will never surpass humans.

Therefore, in order to implement high-level-computers, that is to say, computers-like-brain, it will be necessary to focus on conscious human capabilities, and how they are impacting the information processing of the system.

DYNAMIC OF CONSCIOUSNESS

Any understanding of consciousness should try to explain a huge set of behaviors associated with consciousness. Chalmers defined some of them (Chalmers, 2013), ranging from apparently "simple" tasks (called third-person data) such as perceptual discrimination of stimuli, integration of different sensory modalities, automatic and voluntary actions, accesses and reportability of internal states, differences between sleep and wakefulness, to phenomena even more difficult to explain (called first-person data), for example perceptual experiences (e.g., the experience of color), bodily experiences (e.g., pain and hunger), mental imagery, emotional experiences, among others. Some useful distinctions to study consciousness also point out the differences between studies of wakefulness and

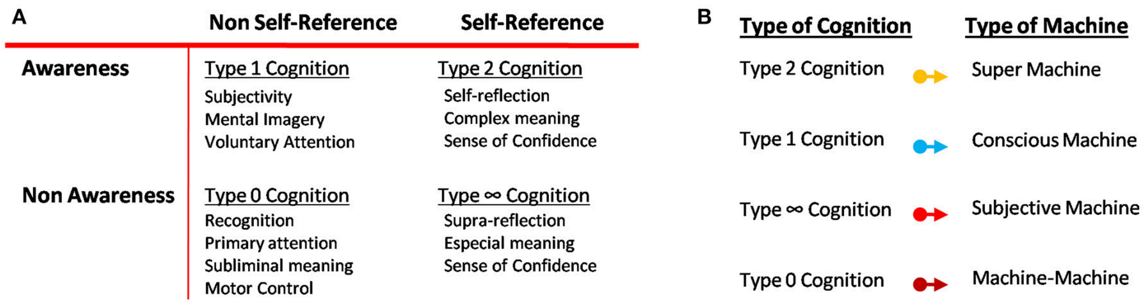


FIGURE 2 | Types of Cognition and Types of Machines. **(A)** Emergent processes related to consciousness and Types of cognition defined from their relations. It is important to highlight that processes associated with moral thought are present in type 1 and type 2 cognition, but not necessarily in the other two types of cognition. **(B)** Types of machines and categories according to different types of cognition, contents, and information processing stated above.

studies of conscious perception or awareness (Chalmers, 2013). The first mechanism would describe the differences between, for example, sleep, vegetative and awake conditions, while the second one tries to explain when and how a perception become consciously perceived, in other words, when we become aware of something (Dehaene et al., 2014). In the end, it is expected that both approaches will help to answer important questions about mechanisms of consciousness, however, these studies do not always include subjective experience, which is assumed to be solved after the understanding of the mechanisms of wakefulness and awareness.

For example, one intriguing characteristic observed from the comparison of subjects in awake condition vs. sleep, vegetative and anesthesia condition is that the neural activity driven by an external stimulation spreads through different areas of the brain when subjects are awake, but remains local when they do not (Rosanova et al., 2012; Casali et al., 2013; Sarasso et al., 2014). Experiments with transcranial magnetic stimulation (TMS) and electroencephalogram recording (EEG) demonstrated this effect. For awake condition, pulses driven by TMS generate richer and sequential EEG signals in different brain areas, and remarkably, the peak of these global activities is lower than in other conditions, where awareness is absent. This signal has been linked with the integration of the brain activity but it is still not clear how integration takes place, which mechanisms allow the global diffusion of each pulse, and why in other than awake condition, the integration remains local.

Additionally, consciousness, awareness and conscious perception, apparently, are not matter of capacity of computation. The brain should not be considered as a computer, neither doing any computation like a computer, as stated above. Although, if someone would like to insist, the brain capacity can be roughly estimated around 20 petaFLOPS, assuming 100 billions of brain cells, 200 firings per second, and 1,000 connections per cell [see other approximations (Martins et al., 2012)], whereas independently of any approximation, 80% of these brain cells (hence its computational capacity) are in the cerebellum, which does not play any important role in conscious perception (Tononi and Koch, 2015). By comparison, the most powerful computer has 93 petaFLOPS [Sunway TaihuLight (Dongarra, 2016; Fu et al., 2016)]. It is however really unlikely that someone ensures that this computer is aware despite its bigger computational capacity. AlphaGo is another example that computational capacity is not the key to improve or reach high-level tasks. The last version AlphaGo zero defeats previous AlphaGo versions but uses less computational resources, suggesting the importance of learning algorithms and neural network architecture to solve complex high-level tasks (Silver et al., 2017).

Nevertheless, evidence has shown that conscious perception needs between 200 to 400 ms (Dehaene and Changeux, 2011) while the processing and integration of information at low-level tasks only need 40 ms. In other words, when we consciously perceive, any processing of information is temporally decreasing between 500 up 1,000%. Experiments, where subjects were exposed to masked stimuli (words or pictures which are masked by previous stimuli), have showed that conscious perception (i.e., subjects report seeing the stimulus) is correlated with a positive peak in Event-related potentials (ERPs) which appear 300–500 ms after the stimulus presentation (**Figure 3A**; Dehaene and Changeux, 2011; Herzog et al., 2016). It is interesting to notice that the neural activity for some cortical regions seems to show a shortly decrease of activity, while other areas showed a later peak around 300–400 ms (Del Cul et al., 2007). This response is called P3b and has not uniquely associated with perception but also with attention and memory processes. The mechanism suggested as an explanation of P3b is a sustained stable activity in recurrent cortical loops. Another mechanism proposed as a marker of conscious perception, called synchrony, has been also observed within a window of 200–400 ms. High-contrast human faces were presented in normal and inverted orientation (Rodriguez et al., 1999), and synchrony was observed around 250 ms each time that faces were recognized. Synchrony was mainly between occipital, parietal and frontal areas (**Figure 3B**). Furthermore, a new pattern of synchrony (in the gamma range) emerged around 720 ms during the motor response. One notable phenomenon from this experiment is the phase scattering presented between these two synchronic responses (Varela et al., 2001). At this time, the probability of finding synchrony between two EEG electrodes was below the level observed before stimulation (**Figure 3B**). This phase scattering and phase synchronization show an interesting kind of alternation or maybe interference, which should be explained by any theory of consciousness.

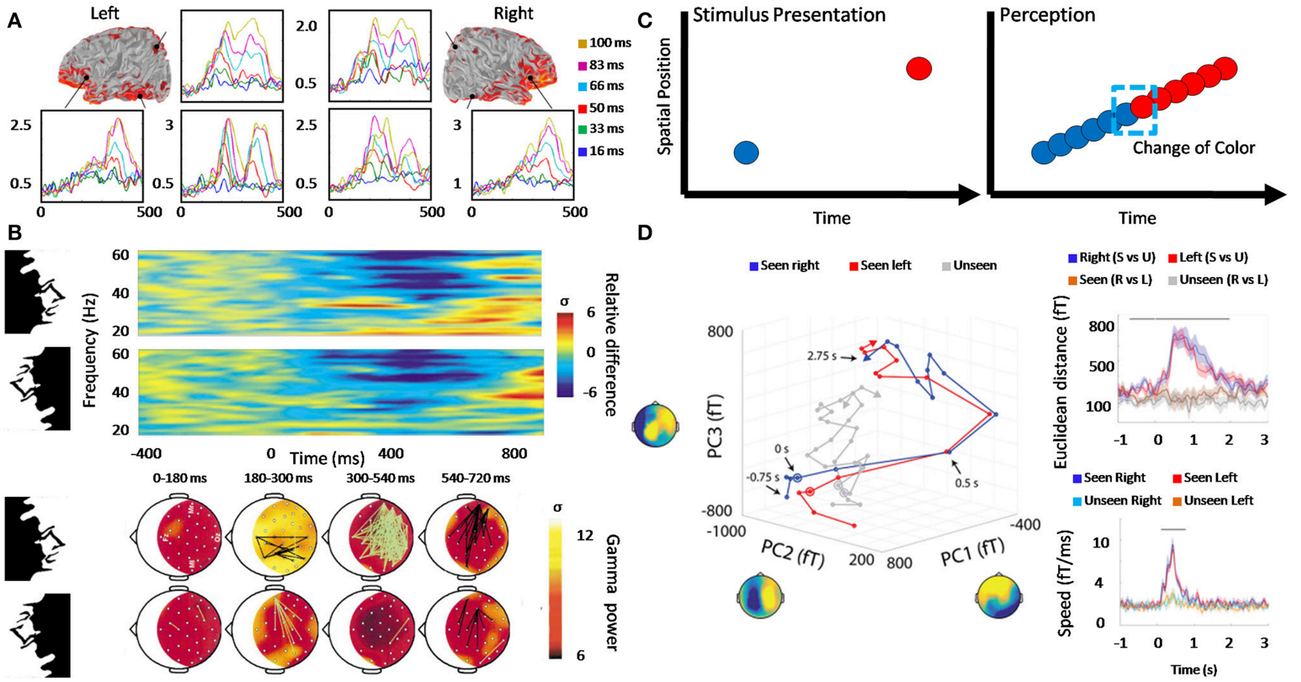


FIGURE 3 | Neural dynamic associated with awareness and some experimental evidence. **(A)** Three cortical areas recorded in left and right hemisphere (posterior parietal, posterior ventral temporal, and inferior frontal) present slightly different types of activity evoked by masked targets. The peak in the condition of maximal visibility is associated with P3b (around 370 ms). Two phases of cortical activation can be recognized, the first previous to 300 ms corresponds to the activity from the occipital pole toward both parietal and ventral temporal sites. The second phase, after 300 ms, is characterized by a high-amplitude activity, which mainly appears in ventral prefrontal cortex together with a re-activation of all previous posterior areas. Colors represent six different conditions where the time of the target-mask stimulus onset asynchrony increased in value, allowing the same stimulus to cross a hypothetical threshold from subliminal processing to conscious perception. Adapted from Del Cul et al. (2007). **(B)** When high contrast faces are presented to normal subjects a long distance synchrony during face-recognition appears around 200 ms at 40 Hz frequency band. Additionally, the effect disappears if the same stimulus is reversed, avoiding the recognition. Another period of synchrony also appears during the motor response and crucially, a transient phase scattering between both synchronic phases showed a decrease in the probability of synchrony. Upper chart is the time-frequency synchrony activity and inferior chart corresponds to the perception condition mapped onto surface electrodes, where black lines indicate a significant level of synchrony, and green lines indicate a marked phase scattering between electrodes. Adapted from Varela et al. (2001) and Rodriguez et al. (1999) with permission of Springer Nature. **(C)** In the color phi phenomenon, two disks are shown at different positions with a rapid succession, inducing the illusion of only one disk which changes the color around the middle trajectory. This phenomenon is contrary to a continuous perceptual dynamic because the observer does not have the opportunity to know in advance the new disk color, especially if the perception is not retrospectively built. Adapted from Herzog et al. (2016). **(D)** Activity trajectories in Principal Component (PC) space of visual conscious perception (red and blue) are different than unconscious perception trajectories (gray). For simplicity, only the first three PCs for subject 2 are shown. The upper-right chart shows the group average Euclidean distance between temporal points for each trajectory [blue right (seen vs. unseen), red (left seen vs. unseen), purple seen (right vs. left), and gray unseen (right vs. left)]. Inferior-right chart corresponds to group average speed of activities trajectories at each time point. Horizontal black lines indicate significant difference (2-way ANOVA, $p < 0.05$, cluster based on permutation test). Adapted from Baria et al. (2017).

A recent experiment has additionally demonstrated a transient neural dynamic during visual conscious perception (Baria et al., 2017), challenging sustained activity mechanisms as broadcasting and integration, and suggesting initial-state-dependent neural dynamics. Neural activity, previous, during and post stimuli, was measured with magnetoencephalography (MEG). Subjects were asked to recognize the direction of Gabor stimulus (left or right) and inform if the stimulus had been consciously perceived (stimuli were manipulated to induce around 50% of conscious perception in each subject). Then, neural activity was divided into different frequency bands to calculate the multi-dimensional state space trajectory computed with principal component analysis (PCA). In the band 0.05–5 Hz, trajectories of conscious (seen) and unconscious (unseen) trials were clearly separable (**Figure 3D**) by Euclidean distance (**Figure 3D** upper right). Crucially, the speed of population activity, measured as a point trajectory in the state space vs. time (ms), showed an acceleration and switch in dynamics after stimulus onset, with a peak around 400 ms (**Figure 3D** inferior right). Moreover, conscious stimuli perception was predicted from the activity up to 1 second before stimulus onset (Baria et al., 2017).

Until now, it is not clear that integration, P3b response and/or synchrony are markers of conscious perception or awareness (Gaillard et al., 2009; Mudrik et al., 2014; Silverstein et al., 2015) and there is no consensus if one exclusive marker can be actually identified. Even so, they can still be markers of "contents" construction at conscious and unconscious level. Most theories about consciousness assume that the construction of contents of consciousness is part of the same phenomenon that they call consciousness, in the sense of awareness. Nevertheless, it is equally reasonable to think that the constructions of contents and awareness are two different dynamics of one process, as transient dynamics suggest, or even two completely different processes. One alternative is to think that the construction of contents is a separated process and previous to the process of becoming aware of these contents. So, we should speak about neural objects, also avoiding "the container" interpretation of consciousness. If this is correct, much recent research on consciousness and conscious perception would be inferring information about the construction of these neural objects that are not necessarily associated in a causal way with consciousness itself. Thus, awareness is one process to explain, and the construction of a perception or objects of consciousness would be another. Integration, P3b and synchrony would be, in this sense, part of the construction of neural objects, but not part of the awareness moment where the object becomes part of our conscious perception. Chronologically, one first stage of information processing should be the constructions of these objects and a second stage would be the awareness of them. These processes would be independent and only from their interactions, as the observer and the observed at the same time, the conscious perception of internal and external neural objects would emerge avoiding the "Cartesian theater" interpretation (Lycan and Dennett, 1993). In other words, it is admissible to be aware without conscious perception of some objects, and "perceive" without awareness about this perception.

Additionally, conscious perception is not always differentiated in awareness and self-reference, but here the distinction is made in order to define clearly different levels of cognition, which would describe two processes of the same conscious phenomenon. In other words, it is possible to state that information processing can be divided into different stages (**Figure 4**), where awareness is related to one of these stages and self-reference with the recursive processing of this stage. The differences between fast time processing for cognition type 0 (∼40 ms) and a slow time processing for type 1 (∼200 ms) have stimulated the idea of Two-Stage Model (Herzog et al., 2016). This is to say that the flux of activity (or inactivity) would need at least two different stages (from which types of cognition emerge), where the first stage corresponds to automatic, non-voluntary control and unconscious information processing, while the second stage would involve a break in this dynamic to allow awareness. Furthermore, it is proposed here that the recursive processing of awareness within the same neural objects will allow the emergence of self-reference process (**Figure 4**).

Other experiments also suggest a discrete mechanism instead of a continuous perception mechanism (VanRullen and Koch, 2003; Chakravarthi and VanRullen, 2012; Herzog et al., 2016). For example, evidence for the discrete mechanism of perception comes from psychophysical experiments where two different stimuli are presented with a short time window between each other. In these experiments, subjects perceived both stimuli as occurring simultaneously, suggesting a discrete temporal window of perception integration (VanRullen and Koch, 2003; Herzog et al., 2016). The most relevant experiment supporting a discrete perception is the color phi phenomenon (**Figure 3C**). In two different locations, two disks of different color are presented in a rapid succession. The observer perceives one disk moving between both positions and changing the color in the middle of the trajectory. Theoretically, the experience of changing color should not be possible before the second disk is seen. Therefore, the perception should be formed retrospectively, which is contrary to continuous theories (Koler and VonGrünau, 1976; Bachmann et al., 2004; Herzog et al., 2016).

Another characteristic is the apparent "interference" between different types of information processed in human conscious behavior. For instance, rational calculations (e.g., resolve a mathematical problem) interfere with kinaesthetic performance (Shea and Frith, 2016). To illustrate, solving a mathematical equation while cycling or dancing at the same time can be practically impossible. This observation suggests that conscious perception would be imposing a balance between different processes. Computational interpretation of this observation will try to explain the interference between different kinds of information as a competition for computational capacity or resources. However, as it is stated above, computational capacity apparently is not playing any crucial role in perception. This analogy also assumes processing of information in a digital way, which could not be the best approach to understand the brain.

Finally, some results from behavioral economics and decision making have shown that cognitive biases are not according to classical probability frameworks (Pothos and Busemeyer, 2013). It means that it is not always possible to describe emergent brain properties with classical and efficient probabilities way. For example, when one tries to explain, for one side, the biological mechanisms in the brain, and on the other, the human psychological behavioral, crucial differences appear. Some research and theories have shown that the dynamics of neural systems can be interpreted in a classic probabilities framework (Pouget et al., 2000; Quiroga and Panzeri, 2009), like good estimator and predictor of external stimuli. While other results, mainly from economic psychology, show cognitive fallacies (Ellsberg, 1961; Gilovich et al., 2002; Moore, 2002; Machina, 2009). These results are incompatible with the classical probability theories (Pothos and Busemeyer, 2013) and can be reconciled only after an extra processing of information in experimental subjects. Therefore, these disconnections between some neural activities in the brain (as classical systems), the emerged human behavior and some of their cognitive capabilities (non-classical systems), and then another possible classical system suggest complex multiple separate systems with interconnected activity (**Figure 4C**). How can some cognitive capabilities, with apparently non-classical dynamic, emerge from apparently classical, or semi-classical systems as neural networks? It is one open question that any theory of consciousness should also try to explain.

AN ALTERNATIVE: CONSCIOUSNESS INTERACTION HYPOTHESES

If consciousness is not a matter of computation capacity, given that temporal efficiency decreases in its presence, it could be due to its architecture. Many theories have tried to explain how

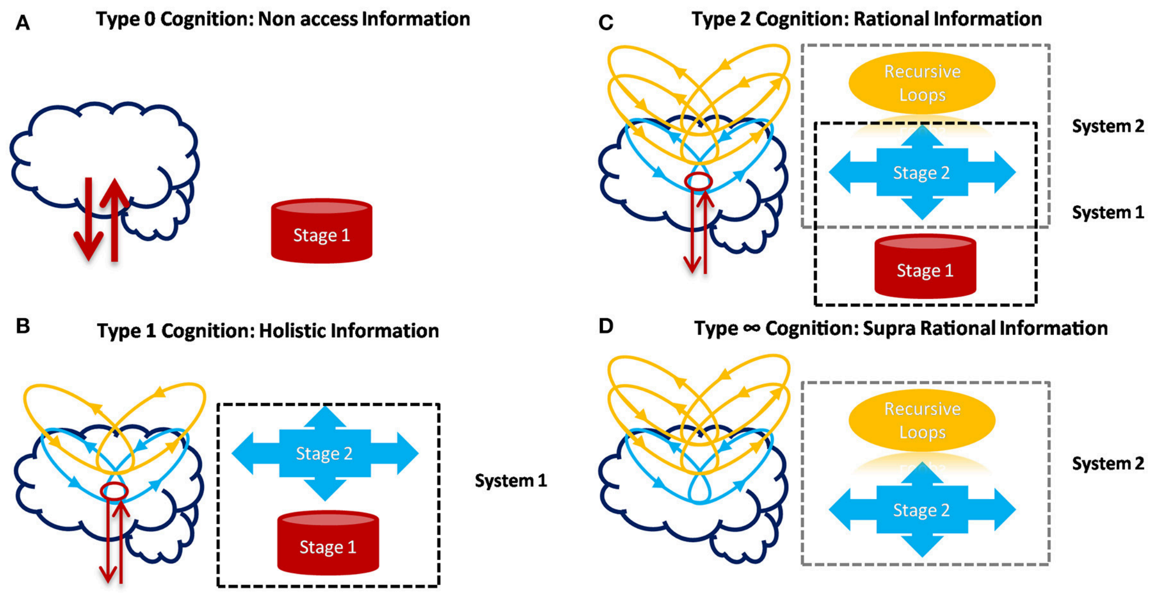

A
Type 0 Cognition: Non access Information
Stage 1
B
Type 1 Cognition: Holistic Information
Stage 2
Stage 1
System 1
C
Type 2 Cognition: Rational Information
Recursive Loops
Stage 2
Stage 1
System 2
System 1
D
Type ∞ Cognition: Supra Rational Information
Recursive Loops
Stage 2
System 2

FIGURE 4 | Types of Cognition, their relation with possible systems and stages of information processing. **(A)** Stage 1 corresponds to automatic and non-conscious processes (classical information) in principal layers. It is associated with Type 0 Cognition. **(B)** Stage 2 is related to awareness and conscious perception as holistic information (Type 1 Cognition) when two or more principal layer interact. Both stages form the non-classical system 1 (linked with psychological features), which is not necessarily deterministic in a classical way. **(C)** Recursive loops of stage 2 would correspond to conscious manipulation processes of contents (Self-reference). From the interaction of stage 2, their recursive loops and re-entry of information with system 1, another classical and deterministic system 2 would emerge. **(D)** The system 2 alone and without interacting with system 1 would correspond to Type ∞ cognition. This type of cognition is a hypothetical/speculative scenario emerged by the relations of Awareness and Self-Reference components in our theoretical framework. However, its existence is doubtful considering that system 2 in living beings, would need system 1 to emerge.

consciousness emerges from the brain (Dehaene et al., 2014; Tononi et al., 2016). However, these theories are incomplete although they might be partially correct. The incompleteness is in part because most of these theories are descriptions of the phenomenon, instead of explanatory theories of the phenomenon. By way of example, Classical Mechanics and Theory of evolution are explanatory theories (although an explanatory and/or complete theory does not ensure that it is correct). Descriptive theories focus on how the phenomenon works, use descriptions without causal mechanisms even when they claim it, and without deductive general principles, i.e., they often start from the object of study to deduce specific/particular principles rather than deducing general principles and in consequence explaining the object of study. Furthermore, incomplete theories do not answer one of these fundamental questions: What is "the object of study"? How does it work? Why? Most commonly, they do not explain "why" something works as it works. In other words, these theories may partially explain and/or describe how consciousness emerges, but they do not explain and do not solve the entire problem. The problem, according to Chalmers (1995, 2013) is to explain both the first-person data related to subjective experience and the third-person data associated with brain processes and behavior. Most of the modern theories of consciousness focus on the third-person data and brain correlates of consciousness without any insight about the subjective experience. Moreover, some of the questions stated above as for example the phase scattering, the transient dynamics, the decrease in the peak of EEG activity driven by TMS, the two stages and two systems division, are not explained, and actually, they are not even well-defined questions that theories of consciousness should explain. Finally, these approaches try to explain awareness and conscious perception in a way that is not clearly replicable or implementable in any sense, neither with biological elements. Some theories also use the implicit idea of computability to explain, for example, conscious contents as the access to certain space of integration; and competition for space of computation in this space, to explain how some processes lose processing capacity when we are conscious.

Another complementary alternative is to understand consciousness as intrinsic property due to the particular form of information processing in the brain. Here, consciousness will be interpreted in this way, as the dynamic interaction/interference (which can be superposition or interference) of different neural networks dynamics, trying to integrate information to solve each particular network problem. More specifically, the brain could be divided into different "principal layers" (topologically speaking, it corresponds to the architecture component) which are also composed by different levels of layers (hypothesis 1), each principal layer as one kind of neural network interconnected at different levels with other networks (**Figure 5**). Each principal layer can process information thanks to oscillatory properties and independently of other principal layers (hypothesis 2); however, when they are activated at the same time to solve independent problems, the interaction generates a kind of

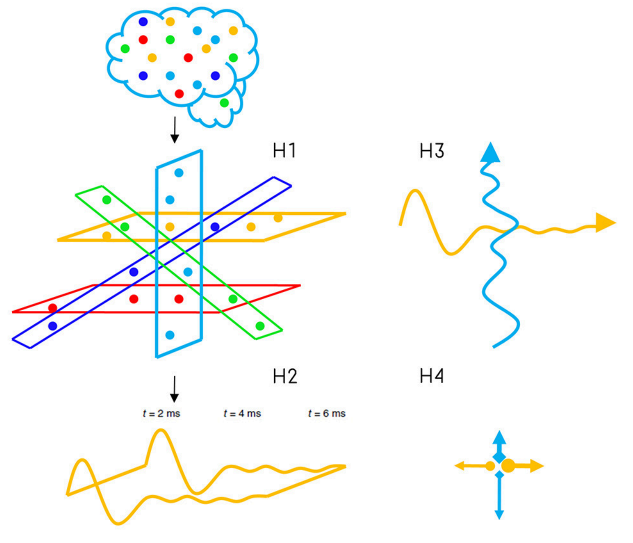


FIGURE 5 | Consciousness Interaction Approach and its four hypotheses. Hypothesis 1: the architecture of the brain can be divided into different independent processing layers (network of networks); hypothesis 2: each layer independently processes information to solve each particular problem; hypothesis 3: two layers activated at the same time can interact/interfere with optimal information processing of each other at different levels, and; hypothesis 4 suggests that this interaction/interference/superposition between at least two principal/independent layers (for example horizontal and vertical) would be associated with the general mechanism of consciousness.

interference on each intrinsic process (hypothesis 3, the processing component). From this interaction and interference would emerge consciousness as a whole (hypothesis 4). I will call it: Consciousness interaction hypotheses. Consciousness would be defined as a process of processes which mainly interferes with neural integration. These processes are an indivisible part of consciousness, and from their interaction/interference, consciousness emerges as a field of electrical, chemical, and kinaesthetic fluctuations.

There are two possible interpretations about these principal layers: the first one is the idea that these principal layers are formed by areas structurally connected, and the second possibility is that they are formed by areas only functionally or virtually connected. In the latter, the functional connectivity should be defined by phases and frequency dynamics to avoid in part the bias about neural activity mentioned above. Experiments and new analyses motivated by these ideas should solve which interpretation is the optimal one. Additionally, the nature of the interference suggested here can sometimes take the form of superposition and other times the form of subtraction in the threshold and/or sub-threshold oscillatory activity associated with neural integration, in two or more principal layers. This interference as a superposition or subtraction would be one possible mechanism to one independent neural process interferes with the other and vice versa (this is not necessarily excitatory and inhibitory neural interactions). Once this interaction has emerged, each principal layer monitors the other without any hierarchical predominance between layers, and if one process disappears, awareness also disappears. In this sense, each principal layer cares about its information processing and the other information processing which can affect them. The oscillatory activity at individual neural layers can be interpreted as one stage (classical information), and when the new activity emerges thanks to interference between principal layers, the second stage would emerge (non-classical information) forming one system. Then, the recursive action of the second stage would allow the emergence of a second system. In the end, both systems as a whole of layers and interactions would be the field of consciousness which cares about its own balance to be able to solve each layer problem.

The idea of "care about something" could also explain in part the subjectivity experience. Each layer cares about some states more than others, based on previous experiences and learning (Cleeremans, 2011), but also grounded on the intrinsic interaction between principal layers defined above, which allow them to solve their information processing problems. In other words, depending on the degree and type of interference for a certain experience, the system would feel one or another feeling, even if the external stimulation (perceptually speaking) is the same for many subjects. The subjectivity, at least preliminarily, would not directly be more or less neural activity. It would be related to the type and degree of interaction between principal layers emerged by learning, balancing processes thanks to plasticity and sub-emergent properties, which all together try to keep the balance of the whole system. This plasticity would be part of emergent and sub-emergent properties of dynamical systems, probably driven by oscillations and neurotransmitters. The system would be trained, first by reinforcement learning and later through also voluntary and conscious learning.

These hypotheses might allow us to replicate some neural activities illustrated above, some features of conscious behavior and to explain, for example, why the brain is not always an efficient machine as it is observed in cognitive fallacies, why decisions are not always optimal, especially in moral dilemmas, why it is possible to observe an apparent decrease in processing capacity between different types of information processing in human conscious behavior when we try to perform rational vs. kinaesthetic tasks. The sustained interference mechanism would break the stability in principal layers triggering different responses in each one, breaking synchrony, local integration and spreading activity and de-activity around principal layers. It could explain in part the transient dynamic, the scattering phase between two synchronic phases associated with conscious perception and motion reportability, or why the activity after TMS in awareness is globally spread, and more interesting, it would allow us to implement a mechanism on other machines than biological machines, if important soft properties and physical principles of brains, as plasticity and oscillations, are correctly implemented in artificial systems. Although these ideas still do not answer the "why" question of a complete theory of consciousness, they are part of a global framework on codification, processing of information, mathematical category and physical theories, which will intent to answer that question and will be developed in further works.

Some important differences of this framework with previous approaches are: (1) awareness would emerge from the property of breaking neural integration, synchrony and symmetry of the system; (2) conscious perception would correspond to dynamics operations between networks more than containers formed by networks in which to put contents. In this sense, consciousness is a distributed phenomenon by essence and the semantic of "neural objects" should be used instead of contents; (3) consciousness would be related to mechanism of oscillatory superposition, interference and sub-emergent properties as oscillatory plasticity; (4) consciousness interaction hypothesis could be an implementable mechanism for artificial intelligence.

Finally, one crucial observation emerges from this discussion. Consciousness interaction hypothesis requires a balance of interaction/interference between different processes involved in its emergence to keep, in fact, the interaction. Otherwise, one principal layer would dominate the interrelated activity, driving the activity in other layers without exchange of roles, which is the opposite approach (during other non-conscious conditions, for example, it could be the case). That is why extraordinary capacities in some processes are compensated with normal or sub-normal capacities in other processes of information when we are conscious.

TYPES OF COGNITION AND TYPES OF MACHINES

Consciousness interaction is a different framework, therefore it is necessary to re-interpret some definitions from previous theories about consciousness (Dehaene et al., 2014). **Conscious states** as different levels of awareness (vegetative, sleep, anesthesia, altered states, aware) would correspond to different types and degrees of interaction or interference between different networks. In this sense, coma patients would miss some crucial interactions between some principal layers which are important for "neural objects" constructions; while during anesthesia, the activity of some principal layers may be only locally affected, losing the optimal balance between layer interactions/interference. In consciousness interaction hypothesis, consciousness is not a particular state neither has possible states; this is a crucial difference regarding common definitions and theories. Consciousness should be interpreted as an operation/process itself. **Contents of consciousness** as elements or information in the external or internal world which at times are part of our conscious perception, would correspond to superposition of different oscillation on certain "intersection points" of interference between networks or the network points (nodes) which are influenced/affected by this interference/interaction (probably in a scattered/sparse way). These "neural objects" can be formed even without awareness. In this case, the neural object is restricted to the universe of one principal layer and their local dynamic. However, they become part of the conscious perception only when two or more principal layers start to share these elements to solve their layer problems. Only at this moment, a neural object appears as part of the field of consciousness. Finally, **conscious processing** is normally defined as the operations applied to these contents/neural objects. In consciousness interaction framework, it would correspond to constants or sustained "loops" of interference/interaction on this "intersection points" and its dynamic evolution (probably through sub-threshold resonant circuits).

With similar definitions (without this particular interference interpretation) and their relations, Shea and Frith have identified four categories of cognition (Shea and Frith, 2016) depending if neural objects and cognitive processes are conscious or not. In previous sections, these four types of cognition were re-defined (**Figures 2**, **4**) from the inter-relation between awareness and self-reference. In summary, Type 0 cognition corresponds to cognitive processes which are not conscious neither in their neural objects nor operations applied to these objects. Type 1 cognition is a set of cognitive processes where neural objects are consciously perceived, however operations on them are not manipulated. Type 2 cognition would correspond to neural objects and operation on these objects consciously perceived and manipulated. Finally, what I have called Type ∞ cognition (Signorelli, 2017) can be understood as cognition without any kind of neural object consciously perceived, but operations on these objects are consciously manipulated. According to these definitions (**Figures 2**, **4**), it is also possible to relate these categories with four categories of machines and their information processing capabilities (Signorelli, 2017): (1) The **Machine-Machine Type 0 Cognition** would correspond to machines and robots that do not show any kind of awareness. These systems cannot know that they know about something that they use to compute and solve problems. Machine-Machine is not intelligent according to the general definition in section A Sub Set of Human Capabilities and their processes are considered low cognitive capabilities in human. Examples are robots that we are making today with a high learning curve. (2) **Conscious-Machine Type 1 Cognition** would have awareness and all the processes of type 1 cognition in humans. This is a very smart machine, however, it cannot control voluntary their inner manipulations even when they can extract meanings of their own "contents." As well as humans, they will show wrong answers to simple questions as for example cognitive fallacy questions, mainly because the system accesses to a wider range of information thanks to first levels of interference/interaction between networks (Holistic information), however, some optimal or specific algorithmic calculations may become intractable. (3) **Super Machine Type 2 Cognition** would be the closest machine to human, at least cognitively speaking. If this machine can reach awareness and self-reference in the sense illustrated here (not only computationally), they should show some kind of "thoughts" associated with consciousness as a whole of rational and emotional processes. In this case, they will have some moral thinking, even when their moral can be completely different than the human moral. The moral thinking is not necessarily restricted to the human morality, because as also happen in different human communities and even human subjects, machines may develop their own type of morality, and this morality can also be non-anthropocentric. Nevertheless, the requirement for any type of moral thinking is the attribution of correct and incorrect behaviors based on what the system cares about the environment, peers and itself, according to a balance between rational and emotional intelligence. If the machine has the ability of awareness and self-reference, they will develop, or they already developed self-reflection, sense of confidence, some kind of empathy among other processes mentioned to reach moral thoughts. In these machines, "contents" are conscious and the cognitive process is deliberate and controlled thanks to a recursive and sustained interference/interaction at certain intersection points from different networks (e.g., reasoning). (4) **Subjective-Machine Type ∞ Cognition** are different than humans, even if they could reach some important features of human intelligence. They are defined according to type ∞ cognition, where awareness is missed but self-reference would still be there. A clear analogy with humans is not stated here, even when the presence of self-reference as a kind of monitoring process without awareness could be reported in humans. However, the hypothesis about this type of machines is related to Supra reasoning information emerged from organization of intelligent parts of this supra system (e.g., Internet), where systems would show some special kind of self-reflection, sense of confidence, even when they will probably not be able to extract meaning of their own "contents," or if they can, it will be especially different than humans.

Some previous works have been also tried to generalize and characterize some features of consciousness and their connection with types of machines and/or artificial systems (Aleksander and Morton, 2008; Wang, 2012). For example in Arsiwalla et al. (2017), even though that article still keeps a computational view of consciousness and social interactions, they conclude that consciousness is not only due to computational capacity and put emphasis in social interactions (which can also be related to emotions) as a trigger of consciousness. Another example is Gamez (2008), where some categories defined can be close to some types of machine mentioned above. However, some crucial differences with these articles are: (1) here, types of machines directly emerge from previous theoretical and experimental definitions of types of cognition. In this context, types of machines are general categories from the definitions of cognition and its relation with consciousness. (2) Additionally, here, it is not assumed any special optimization processes to achieve consciousness, actually quite the contrary, interference processes as non-optimal processes and some still missing properties of soft materials/brains would be associated with its emergence.

Due to these non-optimal processes, each type of machines has limitations (Signorelli, 2017, 2018). For instance, conscious machine type 1 cognition will reach consciousness but it does not have strong algorithmic calculation capabilities or rational/logical intelligence, because accuracy is lost in favor of consciousness as fast access to holistic information. Subjective machines type ∞ cognition probably will not be able to interact physically with us, and even less dance like us or feel like us, however, it is the most likely scenario where machines and computers would overtake some humans capabilities, keeping the current hardware in a non-anthropomorphic form. For this machine, the subjective experience could be something completely different to what it means for humans. In other words, Subjective Machines are free of human criteria of subjectivity. Eventually, Super Machine is the only chance for AI to reach and exceed human abilities as such. This machine would have subjective experiences like humans, at the same time that it would have the option to manipulate the accuracy of its own logic/rational process; however, it is also vulnerable to what subjective experiences imply: the impact of emotions in its performance and biased behavior as humans.

IMPLICATIONS FOR ARTIFICIAL INTELLIGENCE AND CONSCIOUS MACHINE

Any attempt to accomplish conscious machines and try to overcome human capabilities should start with some of the definitions stated previously. First, it is necessary to define a set or subset of human capabilities which are desirable to imitate or even exceed. This is, actually, a common approach, the only difference is the kind of features which have been replicated or attempted to replicate. According to this work, most of them are still low-level cognitive tasks for brains. Also in this article, the subset can be considered a very ambitious group of characteristic: Autonomy, reproduction and moral. Autonomy is already one characteristic considered in AI. Research is currently working to obtain autonomous robots and machines, and nothing opposes to the idea that eventually an autonomous robot can be created. It would probably not be autonomous in the biological sense, but it could reach a high-level of autonomy. The same can be expected for reproduction. Machine reproduction will not be a reproduction as in biological entities, but if robots can repair themselves and even make their own replications, the reproduction issue can be considered reached, at least functionally speaking. However, it is not obvious that genuine moral thinking can be achieved by only improving computational capability or even learning algorithms, specifically, if AI does not add something which is an essential part of the human being: consciousness.

Moreover, when some characteristics of human brains are critically reviewed, consciousness is identified as an emergent property that requires at least two other emergent processes: awareness and self-reference. Thanks to these processes, among others, it is expected to develop high-level cognition which involves processes as self-reflection, mental imagery, subjectivity, sense of confidence, etc, which are needed to show moral thinking. In other words, the way to reach and overcome human features is trying to implement consciousness in robots to attain moral thinking.

However, to try to implement consciousness in robots, a theory is needed that can explain, biologically and physically speaking, consciousness in human brains, dynamics of possible correlates of consciousness, the psychological phenomenon associated with conscious behavior and at the same time, explore mechanisms which can be replicated into machines. It should not be mere descriptions of which areas of the brain are activated or which are the architectures of consciousness, if the interaction between them, from which consciousness would emerge, is not understood. Therefore, the understanding of emergent properties is not enough and the consideration of crucial plasticity properties of the soft materials in biology, as oscillations, stochasticity, and even noise are very important to also understand sub-emergent properties as plasticity changes influenced by voluntary or conscious activity. For one side, a more complete theory of consciousness is needed, which relates complex behavior with physical substrates and for another side, we need neuromorphic technologies to implement these theories.

One of the main attempts of this paper was to show a possible structure for consciousness, founded on a non-intuitive kind of interaction: oscillatory superposition and interference between networks of networks defined as structural and/or functional organizations changing dynamically. These principal networks try to solve particular problems, and when all of them are activated, sharing and interfering on their own oscillatory processes as a whole, the field of consciousness would emerge as a process of processes. Additionally, another main attempt explored here was to make evident some paradoxical consequences of trying to reach human capabilities. Thus, types of cognitions were defined not only to show different conscious processes, but also to show that from these categories, it is possible to define four types of machines regarding the implementation of consciousness into machines, and their limitations.

For example, if we can reach the gap to make conscious machine type 1 or 2 cognition, these machines will lose the meaningful characteristics of being a computer, that is to say: to solve problems with accuracy, speed and obedience. Any conscious machine is not a useful machine anymore; unless they want to collaborate with us. It means the machine can do whatever it wants; it has the power to do it and the intention to do it. It could be considered a biological new species, more than a machine or only computer. More important: according to our previous sections and empirical evidence from psychology and neuroscience (Haladjian and Montemayor, 2016; Signorelli, 2017), it is not possible to expect an algorithm to control the process of emergence of consciousness in this kind of machines, and in consequence, we would not be able to control them. In other words, even if it were possible to replicate consciousness and high-level cognition, each machine would be different to the other in a way that we are not going to control. If someone expects to have a super-efficient machine, it would be quite the contrary, each machine would be a lottery just as it is when people meet each other.

With this in mind, three paradoxes appear. The first paradox is that the only way to reach conscious machines and potentially overcome human capabilities with computers is by making machines which are not computers anymore. If it is considered that a subset of main features on machines is the capacity to be accurate and fast solving problems, from comments above, any system with subjective capabilities is not accurate anymore, because if they replicate high-level cognitions of human, it is also expected that they will replicate the experience of color or even pain, in a way that it will also interfere with rational and optimal calculations, as well as in humans. The second paradox is that when we make conscious machines type 1 and/or type 2 cognition, a process of interference, due to consciousness, will affect the global processing of information, allowing extraordinary rational or emotional abilities, but never both extraordinary capabilities at the same time or even in the same individual, due in part to how the intrinsic and non-controlled emergent processes associated with consciousness would work. In fact, if the machine is a computer-like-brain, this system will require a human-like-intelligence that apparently also requires a balance between different intelligence, as stated above. Hence, machines type 1 or type 2 cognition would never surpass human abilities, or if it does, it will have some limitations like humans. The last paradox, if humans are able to build a conscious machine that overcomes human capabilities: Is the machine more intelligent than humans or are humans still more intelligent because we could build it? The intelligence definition would move again, according to AI successes and new technologies reached.

The ultimate goal of all these discussions is to emphasize that trying to make conscious machines or trying to overcome humans is not the path to improve machines, and indeed, to overcome humans is a contradiction in itself. Futurists speak about super machines with super-human characteristics, but they stimulate these ideas without any care about what means to be a human or even simple, but amazing kind of animals which are still much smarter than computers. To make better machines, science should not focus on anthropocentric presumptions nor compare the intelligence of a machine with human intelligence. The comparison should be according to a general definition of intelligence, as it is stated above. This definition is complex enough and very ambitious goal for any kind of AI. In this way, better machines will be the type 0 and ∞ cognition without anthropomorphic requirements, which will be able to find different solutions to human problems and probably unimaginably better than humans. These machines would be able to imitate some human behavior if needed, but never achieve the genuine social or emotional interaction that humans and animals already have.

On another side, the question about replicating human capabilities is still interesting and important, but for reasons which are not efficient, optimal or better machines. The interest of studying how to implement genuine human features in machines is one academic and even ethical goal, as for example a strategy to avoid animal experimentation. As it was shown above, robots and machines would not be able to replicate the subset of the human being if they do not replicate important features of brains-hardware mentioned previously. These properties are apparently closely connected with important emergent properties which are a fundamental part of consciousness, and some features of consciousness are needed to replicate moral thinking as a crucial and remarkable capability of human beings. That is why, to really understand the biological complexity and mechanisms associated with these emergent properties, the construction of artificial machines based on soft and biological properties/principles can allow us to manipulate and find different kinds of mechanisms until reaching some of the interesting characteristics of living beings. This approach will not take us to more efficient machines, quite the contrary, these machines will be inefficient and if, for instance, type 1 cognition is achieved, they will be closer to some animals, more than good and simple current machines.

That is why, finally, AI could be divided in (1) Biological-Academic Approach, to achieve human intelligence for academic proposes, as for example, instead of using animals to understand consciousness, trying to use robots to implement theories about how consciousness or other important biological features are working. However, once the ultimate goal is reached, for instance, the understanding of consciousness, the knowledge should not be used to replicate or massively produce conscious machines. It would be essentially an ethical question, at the same level or even more intractable than cloning animal issues. (2) Efficient Approach, to make better robots and machines, which can help us with important tasks that are difficult to perform or improve the human performance. The goal is efficiency and performance. In this approach, some principles from biology can be useful, such as modern applications of neural networks, but the final goal would not be to achieve high-level cognition. The implementation in silicon of biological and physical principles of high-level cognition in humans and animals will help us to improve some performances, but these technologies will never replicate truly social interactions, and it should not be expected, because these kinds of interactions are apparently connected with hardware dependences of biological brains. Of course, it is expected to imitate some of them and even incorporate mixed systems between efficient silicon architectures and inefficient soft materials to reach this goal, but any attempt should be conscious of their intrinsic limitations.

CONCLUSIONS

These comments seek to motivate discussion. The first objective was to show typical assumptions and misconceptions when we speak about AI and brains. Perhaps, in sight of some readers, this article is also based on misunderstandings, which would be another evidence of the imperative need for close interaction between biological sciences, such as neuroscience, and computational sciences. The second objective was tried to overcome these assumptions and explore a hypothetical framework to allow conscious machines. However, from this idea emerge paradoxical conclusions of what a conscious machine is and what it implies.

The hypotheses stated above are part of a "proof of concept" to be commented and reformulated. They are part of a work in progress. Thanks to category theory, process theories and others theoretical frameworks, it is expected to develop these ideas on consciousness interaction hypothesis more deeply and relate them with other theories on consciousness, its differences and similarities. In this respect, it is reasonable to consider that a new focus that integrates different theories is needed. This article is just the starting point of a global framework on the foundation of computation, which expects to understand and connect physical properties of the brain with its emergent properties in a replicable and implementable way to AI.

In conclusion, one suggestion of this paper is to interpret the idea of information processing carefully, perhaps in a new way and in opposition to the usual computational meaning of this term, specifically in biological science. Further discussions which expand this and other future concepts are more likely to be fruitful than mere ideas of digital information processing in the brain. Additionally, although this work explicitly denies the analogy brain-digital-computer, it is still admissible a machine-like-brain, where consciousness interaction could be an alternative to implement high intelligence in machines and robots, knowing the limitations of this approach. Even if this alternative is neither deterministic nor controlled, and presents many ethical questions, it is one alternative that might allow us to implement a mechanism for a conscious machine, at least theoretically. If this hypothesis is correct and it is possible to reach the gap of its implementation, any machine with consciousness based on brain dynamics may have high cognitive properties. However, some type of intelligence would be more developed than others, because, by definition, its information processing would also be similar to brains which have these restrictions. Finally, these machines would paradoxically be autonomous in the most human sense of this concept.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was partially supported by Comisión Nacional de Investigación Ciencia y Tecnología (CONICYT) through Programa Formacion de Capital Avanzado (PFCHA), Doctoral scholarship Becas Chile: CONICYT PFCHA/DOCTORADO BECAS CHILE/2016 - 72170507.

ACKNOWLEDGMENTS

The author appreciates valuable comments of Marlène Morvan, Peter Williams and reviewers, which helped to improve the quality of some discussions. Additionally, the author would like to express his gratitude to journals, open access and creative commons attribution license, which allowed the partial reproduction of figures and descriptions, while others journals like SAGE publications refused reiteratively grant permission and define their policies of fair use, even regarding a clearly adapted figure. The author strongly believes that these practices should be denounced in favor of fair use of scientific production and open access, in view of a more collaborative environment for the science of the present and future. Finally, the author thanks Frontiers platform and editors to be aligned with these values, support young scientist in different steps of the publication process and contribute to more accessible science for everyone.

REFERENCES

Aleksander, I., and Morton, H. (2008). Computational studies of consciousness. *Prog. Brain Res.* 168, 77–93. doi: 10.1016/S0079-6123(07)68007-8

Alvarez-maubecin, V., Garc,i, F., Williams, J. T., and Bockstaele, E. J., Van. (2000). Functional coupling between neurons and glia. *J. Neurosci.* 20, 4091–4098. doi: 10.1523/JNEUROSCI.20-11-04091.2000

Arsiwalla, X. D., Moulin-Frier, C., Herreros, I., Sanchez-Fibla, M., and Verschure, P. (2017). The morphospace of consciousness. arxiv[preprint] *ArXiv*:20.

Arsiwalla, X. D., Signorelli, C. M., Puigbo, J., Freire, I. T., and Verschure, P. (2018). "Are brains computers, emulators or simulators?" in *Living Machines: Conference on Biomimetic and Biohybrid Systems* (Paris). doi: 10.1007/978-3-319-95972-6_3

Atasoy, S., Donnelly, I., and Pearson, J. (2016). Human brain networks function in connectome-specific harmonic waves. *Nat. Commun.* 7:10340. doi: 10.1038/ncomms10340

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Prog. Brain Res.* 150, 45–53. doi: 10.1016/S0079-6123(05)50004-9

Bachmann, T., Poder, E., and Luiga, I. (2004). Illusory reversal of temporal order: the bias to report a dimmer stimulus as the first. *Vision Res.* 44, 241–246. doi: 10.1016/j.visres.2003.10.012

Baria, A. T., Maniscalco, B., and He, B. J. (2017). Initial-state-dependent, robust, transient neural dynamics encode conscious visual perception. *PLoS Comput. Biol.* 13, 1–29. doi: 10.1371/journal.pcbi.1005806

Barron, A. B., and Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4900–4908. doi: 10.1073/pnas.1520084113

Bekinschtein, T., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *PNAS* 106, 1672–1677. doi: 10.1073/pnas.0809667106

Brefczynski-Lewis, J., A, Lutz, A., Schaefer, H. S., Levinson, D. B., and Davidson, R. J. (2007). Neural correlates of attentional expertise in long-term meditation practitioners. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11483–11488. doi: 10.1073/pnas.0606552104

Bringsjord, S., Licato, J., Sundar, N., Rikhiya, G., and Atriya, G. (2015). "Real robots that pass human tests of self-consciousness," in *Proceeding of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (Kobe), 498–504.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Bullock, T. H., Bennett, M. V. L., Johnston, D., Josephson, R., Marder, E., and Fields, R. D. (2005). The neuron doctrine, Redux. *Science* 310, 791–793. doi: 10.1126/science.1114394

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929. doi: 10.1126/science.1099745

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., et al. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Struct Funct.* 217, 783–796. doi: 10.1007/s00429-012-0380-y

Caporale, N., and Dan, Y. (2008). Spike timing-dependent plasticity: a Hebbian learning rule. *Annu. Rev. Neurosci.* 31, 25–46. doi: 10.1146/annurev.neuro.31.060407.125639

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5:198ra105. doi: 10.1126/scitranslmed.3006294

Chakravarthi, R., and VanRullen, R. (2012). Conscious updating is a rhythmic process. *Proc. Natl. Acad Sci.* 109, 10599–10604. doi: 10.1073/pnas.1121622109

Chalmers, D. (1995). The puzzle of conscious experience. *Sci. Am.* 273, 80–86.

Chalmers, D. (2013). How can we construct a science of consciousness? *Ann. N. Y. Acad. Sci.* 1303, 25–35. doi: 10.1111/nyas.12166

Chappell, J., and Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. *Int J Unconvent Comput.* 3, 211–239. Available online at: https://www.cs.bham.ac.uk/research/projects/cogaff/chappell-sloman-ijuc-07.pdf

Christian, Szegedy, Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2014). Intriguing properties of neural networks. *ArXiv*:1–10.

Cleeremans, A. (2011). The radical plasticity thesis: how the brain learns to be conscious. *Front. Psychol.* 2:86. doi: 10.3389/fpsyg.2011.00086

Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. *Curr. Opin. Neurobiol.* 25, 76–84. doi: 10.1016/j.conb.2013.12.005

Dehaene, S., Lau, H., Kouider, S., Silver, D., Huang, A., Maddison, C. J., et al. (2017). What is consciousness, and could machines have it? *Science* 358, 484–489. doi: 10.1126/science.aan8871

Del Cul, A., Baillet, S., and Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol.* 5, 2408–2423. doi: 10.1371/journal.pbio.0050260

Dongarra, J. (2016). *Report on the Sunway TaihuLight System.* Tech Report UT-EECS-16-742. Available online at: http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Q. J. Econom.* 75, 643–669. doi: 10.2307/1884324

Epstein, R. (2016). *The empty brain*. *Aeon*. Available online at https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Fleischaker, G. R. (1992). Questions concerning the ontology of autopoiesis and the limits of its utility. *Int. J. Gen. Syst.* 21, 131–141. doi: 10.1080/03081079208945065

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., and Rees, G. (2012). Relating introspective accuracy to individual differences in brain structure. *Science* 329, 1541–1544. doi: 10.1126/science.1191883

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Frank, A. (2017). *Minding matter*. *Aeon*. Available online at: https://aeon.co/essays/materialism-alone-cannot-explain-the-riddle-of-consciousness

Fu, H., Liao, J., Yang, J., Wang, L., Song, Z., Huang, X., et al. (2016). The Sunway TaihuLight supercomputer: system and applications. *Sci. China Inform. Sci.* 59:72001. doi: 10.1007/s11432-016-5588-7

Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., et al. (2009). Converging intracranial markers of conscious access. *PLoS Biol.* 7:e1000061. doi: 10.1371/journal.pbio.1000061

Gallistel, C. R., and Balsam, P. D. (2014). Time to rethink the neural mechanisms of learning and memory. *Neurobiol. Learn. Mem.* 108C, 136–144. doi: 10.1016/j.nlm.2013.11.019

Gamez, D. (2008). Progress in machine consciousness. *Conscious. Cogn.* 17, 887–910. doi: 10.1016/j.concog.2007.04.005

Gardner, H. (1999). *Intelligence Reframed: Multiple Intelligences for the 21st Century*. New York, NY: Basic Book.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychol. Sci.* 385–390. doi: 10.1111/j.1467-9280.1993.tb00586.x

Gerstner, W., Sprekeler, H., and Deco, G. (2012). Theory and simulation in neuroscience. *Science* 338, 60–65. doi: 10.1126/science.1227356

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511808098

Goguen, J. A., and Varela, F. J. (1979). Systems and distinctions; duality and complement arity. *Int. J. Gen. Syst.* 5, 31–43. doi: 10.1080/03081077908960886

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Adv. Comput.* 6, 31–88.

Gosseries, O., Di, H., Laureys, S., and Boly, M. (2014). Measuring consciousness in severely damaged brains. *Annu. Rev. Neurosci.* 37, 457–478. doi: 10.1146/annurev-neuro-062012-170339

Haladjian, H. H., and Montemayor, C. (2016). Artificial consciousness and the consciousness-attention dissociation. *Conscious. Cogn.* 45, 210–225. doi: 10.1016/j.concog.2016.08.011

Hebb, D. (1949). *The Organization of Behavior; a Neuropsychological Theory.* New York, NY: Wiley.

Hegel, G. W. F. (2001). Philosophy of Right. *Transition* (Vol. 1). Kitchener, ON: Batoche Books Limited. Available online at: http://dhspriory.org/kenny/PhilTexts/Hegel/PhilosophyRight.pdf

Herzog, M. H., Kammer, T., and Scharnowski, F. (2016). Time slices: what is the duration of a percept? *PLoS Biol.* 14:e1002433. doi: 10.1371/journal.pbio.1002433

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Jonas, E., and Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Comput. Biol.* 13:e1005268. doi: 10.1371/journal.pcbi.1005268

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Kant, I. (1785). *Fundamental Principles of the Metaphysic of Morals*, 1949th ed. New York, NY: L. A. Press, Ed.

Kauffman, L. H. (1987). Self-reference and recursive forms. *J. Soc. Biol. Syst.* 10, 53–72. doi: 10.1016/0140-1750(87)90034-0

Kauffman, L. H., and Varela, F. J. (1980). Form dynamics. *J. Soc. Biol. Syst.*3, 171–206. doi: 10.1016/0140-1750(80)90008-1

Koler, P. A., and VonGrünau, M. (1976). Shape and color in apparent motion. *Vision Res.* 16, 329–335. doi: 10.1016/0042-6989(76)90192-9

Landauer, R. (1999). Information is a physical entity. *Phys. A* 263, 63–67. doi: 10.1016/S0378-4371(98)00513-5

Llinas, R., Ribary, U., Contreras, D., and Pedroarena, C. (1998). The neuronal basis for consciousness. *Philos. Trans. R. Soc. Lond. B* 353, 1841–1849. doi: 10.1098/rstb.1998.0336

Lutz, A., Greischar, L., Rawlings, N., Ricard, M., and Davidson, R. (2004). Long-term meditators self-induce high-amplitude gamma synchrony during mental practice. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16369–16373. doi: 10.1073/pnas.0407401101

Lycan, W. G., and Dennett, D. C. (1993). Consciousness Explained. *Philos. Rev.* 102:424. doi: 10.2307/2185913

Machina, M. (2009). Risk, ambiguity, and the rank-dependence axioms. *Am. Econ. Rev.* 99, 385–392. doi: 10.1257/aer.99.1.385

Martinez-Miranda, J., and Aldea, A. (2005). Emotions in human and artificial intelligence. *Comput. Hum. Behav.* 21, 323–341. doi: 10.1016/j.chb.2004.02.010

Martins, N., Erlhagen, W., and Freitas, R. (2012). Non-destructive whole-brain monitoring using nanorobots: neural electrical data rate requirements. *Int. J. Mach. Consci.* 4, 109–140. doi: 10.1142/S1793843012400069

Maturana, H., and Varela, F. (1998). *De máquinas y seres vivos (Quinta edi)*. Santiago de Chile: Editorial Universitaria, S.A. Available online at: https://antropologiafractal.files.wordpress.com/2015/08/de-mc3a1quinas-y-seres-vivos-autopoiesis-la-organizacic3b3n-de-lo-vivo.pdf

Moore, D. (2002). Measuring new types of question-order effects: Additive and subtractive. *Public Opin. Quart.* 66, 80–91. doi: 10.1086/338631

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proc. IEEE* 86, 82–85. doi: 10.1109/JPROC.1998.658762

Moravec, H. P. (1988). *Mind Children : the Future of Robot and Human Intelligence.* Cambridge, MA: Harvard University Press.

Mudrik, L., Faivre, N., and Koch, C. (2014). Information integration without awareness. *Trends Cognit Sci.* 18, 488–496. doi: 10.1016/j.tics.2014.04.009

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence.* Cambridge University Press. Available online at: https://ai.stanford.edu/~nilsson/QAI/qai.pdf

Nusbaum, M. P., Blitz, D. M., Swensen, A. M., Wood, D., and Marder, E. (2001). The roles of co-transmission in neural network modulation. *Trends Neurosci.* 24, 146–154. doi: 10.1016/S0166-2236(00)01723-9

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., and Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science* 313:1402. doi: 10.1126/science.1130197

Pothos, E. M., and Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behav. Brain Sci.* 36, 255–274. doi: 10.1017/S0140525X12001525

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132. doi: 10.1038/35039062

Quiroga, R. Q., and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci.* 10, 173–185. doi: 10.1038/nrn2578

Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B., and Varela, F. J. (1999). Perception's shadow: Long-distance synchronization of human brain activity. *Nature* 397, 430–433.

Ronald, C., and Nicolelis, M. A. L. (2015). *The Relativistic Brain: How It Works and Why It Cannot by Simulated by a Turing Machine*. Natal: Kios Press.

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M. A., et al. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain* 135, 1308–1320. doi: 10.1093/brain/awr340

Sarasso, S., Rosanova, M., Casali, A. G., Casarotto, S., Fecchio, M., Boly, M., et al. (2014). Quantifying cortical EEG responses to TMS in (Un)consciousness. *Clinical E. E. G. Neurosci.* 45, 40–49. doi: 10.1177/1550059413513723

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–457. doi: 10.1017/S0140525X00005756

Searle, J. R. (1990). Is the brain a digital computer? *Proc. Addres. Am. Philo. Assoc.* 64, 21–37. doi: 10.2307/3130074

Shea, N., and Frith, C. D. (2016). Dual-process theories and consciousness: the case for "Type Zero" cognition. *Neurosci. Consci.* 2016:niw005. doi.org/10.1093/nc/niw005

Signorelli, C. M. (2017). "Types of cognition and its implications for future high-level cognitive machines," in *AAAI Spring Symposium Series* (Berkeley, CA). Available online at: http://aaai.org/ocs/index.php/SSS/SSS17/paper/view/\penalty-\@M15310

Signorelli, C. M. (2018). "Can computers overcome humans ? consciousness interaction and its implications," in *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing* (Berkeley, CA).

Signorelli, C. M., and Arsiwalla, X. D. (2018). "Moral Dilemmas for Artificial Intelligence: a position paper on an application of Compositional Quantum Cognition," in *Quantum Interaction. QI 2018. Lecture Notes in Computer Science* (Nice).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Silverstein, B. H., Snodgrass, M., Shevrin, H., and Kushwaha, R. (2015). P3b, consciousness, and complex unconscious processing. *Cortexv* 73, 216–227. doi: 10.1016/j.cortex.2015.09.004

Simpson, J. R., Snyder, A. Z., Gusnard, D., A., and Raichle, M. E. (2001). Emotion-induced changes in human medial prefrontal cortex: I. During cognitive task performance. *Proc. Natl. Acad. Sci. U. S. A.* 98, 683–687. doi: 10.1073/pnas.98.2.683

Sloman, A. (2007). "Why some machines may need qualia and how they can have them : including a demanding new turing test for robot philosophers," in *Invited presentation for AAAI Fall Symposium 2007*. Available online at: http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#717

Smith, J. D. (2009). The study of animal metacognition. *Trends Cognit. Sci.* 13, 389–396. doi: 10.1016/j.tics.2009.06.009

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *Am. Psychol.* 52, 1030–1037. doi: 10.1037/0003-066X.52.10.1030

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., et al. (2016). "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford, C. A. Retrieved from http://ai100.stanford.edu/2016-report

Stuart, G. J., Dodt, H. U., and Sakmann, B. (1993). Patch-clamp recordings from the soma and dendrites of neurons in brain slices using infrared video microscopy. *Pflug. Archv. Eur. J. Physiol.* 423, 511–518. doi: 10.1007/BF00374949

Tetzlaff, C., Kolodziejski, C., Markelic, I., and Wörgötter, F. (2012). Time scales of memory, learning, and plasticity. *Biol. Cybernet.* 106, 715–26. doi: 10.1007/s00422-012-0529-z

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44

Tononi, G., and Koch, C. (2008). The neural correlates of consciousness: an update. *Ann. N. Y. Acad. Sci.* 1124, 239–261. doi: 10.1196/annals.1440.004

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. Lond. B* 370:20140167. doi: 10.1098/rstb.2014.0167

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Uhlhaas, P. J., Roux, F., Rodriguez, E., Rotarska-Jagiela, A., and Singer, W. (2010). Neural synchrony and the development of cortical networks. *Trends Cognit. Sci.* 14, 72–80. doi: 10.1016/j.tics.2009.12.002

VanRullen, R., and Koch, C. (2003). Is perception discrete or continuous? *Trends Cognit. Sci.* 7, 207–213. doi: 10.1016/S1364-6613(03)00095-0

Varela, F., Lachaux, J., Rodriguez, E., and Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239. doi: 10.1038/35067550

Varela, F. J. (1975). A calculus for self-reference. *Int. J. Gen. Syst.* 2, 5–24.

Varela, F. J., and Goguen, J. A. (1978). The arithmetic of closure. *J. Cybernet.* 8, 291–324. doi: 10.1080/01969727808927587

Wang, Y. (2012). The cognitive mechanisms and formal models of consciousness. *Int. J. Cognit. Inform. Nat. Intel.* 6, 23–40. doi: 10.4018/jcini.2012040102

Conflict of Interest Statement: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership