THE MECHANISMS UNDERLYING THE HUMAN MINIMAL SELF

EDITED BY: Verena V. Hafner, Bernhard Hommel, Ezgi Kayhan, Dongheui Lee,
Markus Paulus and Stephan Alexander Verschoor

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-660-4
DOI 10.3389/978-2-88976-660-4

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

THE MECHANISMS UNDERLYING THE HUMAN MINIMAL SELF

Topic Editors:
Verena V. Hafner, Humboldt University of Berlin, Germany
Bernhard Hommel, University Hospital Carl Gustav Carus, Germany
Ezgi Kayhan, University of Potsdam, Germany
Dongheui Lee, Vienna University of Technology, Austria
Markus Paulus, Ludwig Maximilian University of Munich, Germany
Stephan Alexander Verschoor, Leiden University, Netherlands

Citation: Hafner, V. V., Hommel, B., Kayhan, E., Lee, D., Paulus, M., Verschoor, S. A., (2022). The Mechanisms Underlying the Human Minimal Self.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-660-4

Table of Contents

frontiers | Frontiers in Psychology

EDITORIAL
published: 01 July 2022
doi: 10.3389/fpsyg.2022.961480

Editorial: The Mechanisms Underlying the Human Minimal Self

Verena Hafner[1], *Bernhard Hommel*[2,3]*, *Ezgi Kayhan*[4], *Dongheui Lee*[5], *Markus Paulus*[6] and *Stephan Verschoor*[7]

[1] Adaptive Systems Group, Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany, [2] Department of Psychology, Shandong Normal University, Jinan, China, [3] University Hospital Carl Gustav Carus, Dresden, Germany, [4] Department of Developmental Psychology, University of Potsdam, Potsdam, Germany, [5] Human-Centered Assistive Robotics, Technical University of Munich, Munich, Germany, [6] Developmental Psychology, Ludwig Maximilians-Universität München, Munich, Germany, [7] University of Bremen, Bremen, Germany

Keywords: agents, self, minimal self, robotics, humanoids, cognition

Editorial on Research Topic

The Mechanisms Underlying the Human Minimal Self

The human self is a particularly colorful concept that occupies a central position in the cognitive and social sciences since their existence: it is the agent that is doing the thinking in Descartes' quest for the validity of human knowledge, the target of religious and political persuasion, the ultimate goal of personal development and therapeutic intervention, and the key factor in attributing legal and ethical responsibility. But what is the self? It is often taken as a given, or at least as a useful fiction (as in legal thinking), but rather little is known about how it works, where it comes from, and what its potential might be. Recently, there has been renewed interest in the so-called minimal self (Gallagher, 2000). According to philosophical views, the minimal self (in contrast to a narrative self or verbalized self-concept) refers to a person's phenomenal experience as an acting and perceiving individual in the here and now. In other words, it describes the pre-reflective representation that emerges from concrete sensorimotor experience. Current research has focused on, the sense of agency and body ownership experiences as two central aspects of the minimal self.

Unfortunately, the psychological basis of the minimal self is not well understood. In fact, there is no truly mechanistic approach that at least tries to capture the processes underlying the minimal self. However, important methodological developments and the availability of novel research techniques (such as virtual reality and humanoid robotics), the dramatic increase of interest in the experimental investigation of the minimal self in the recent years, and the convergence of two lines of cognitive theorizing may make the time ripe for the next major step in understanding the minimal self.

One of these lines refers to the concept of embodied cognition. There is increasing dissatisfaction with the idea that human cognition is abstract, symbolic, and entirely disembodied. This dissatisfaction has stimulated approaches that emphasize the role of people's active sensorimotor experience in creating knowledge, including knowledge about oneself. While these approaches still lack mechanistic detail (Hommel, 2016), they raise the possibility that the self is not just a given but something that emerges through experience and learning. This in turn implies that we can study and reconstruct this emergence in developmental experiments and create experimental manipulations that provide causal tests of theories by changing self-representation in predicted ways.

The other line of theorizing that provides important tools for unraveling the mechanisms underlying the self relates to ideomotor theory (Hommel, 2017). Ideomotor theory seeks to identify the mechanisms underlying goal-directed action and, given the assumed role of sensorimotor experience in creating self-representations, these action mechanisms might also contribute to

OPEN ACCESS

Edited and reviewed by:
Eirini Mavritsaki,
Birmingham City University,
United Kingdom

***Correspondence:**
Bernhard Hommel
bh@bhommel.onmicrosoft.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

***Received:** 04 June 2022*
***Accepted:** 21 June 2022*
***Published:** 01 July 2022*

Citation:
Hafner V, Hommel B, Kayhan E, Lee D, Paulus M and Verschoor S (2022) Editorial: The Mechanisms Underlying the Human Minimal Self. Front. Psychol. 13:961480. doi: 10.3389/fpsyg.2022.961480

understanding the mechanisms of self-creation. Unraveling these mechanisms allows researchers to reconstruct selves in artificial agents (Hafner et al., 2020), which provides a very promising testbed for empirical theories of the self. Indeed, the field of cognitive robotics becomes increasingly interested in sensorimotor learning and the re-enactment of sensorimotor experience (Vernon et al., 2015; Schillaci et al., 2016a). Internal models and mechanisms for internal simulation of sensorimotor activity have been found to be promising tools for the implementation of basic cognitive skills in artificial agents (Schillaci et al., 2016b). In particular, the ideomotor idea that decision-making is based on the anticipation of action effects plays a central role in predictive-coding approaches to both artificial agents and humans (Kilner et al., 2015).

Empirical studies of the self also strongly benefit from recent methodological developments in various fields. The study of self-development was stimulated by the availability of non-invasive brain imaging techniques (e.g., Saby and Marshall, 2012), computer-based looking time paradigms and the fine-grained analysis of eye movements and pupil size (e.g., Gredebäck et al., 2010). These methods are supposed to allow the analysis of cognitive mechanisms even in infants, yet are also under debate (Paulus, 2022). Converging ideas in developmental psychology and cognitive robotics have created a new interdisciplinary research area called "developmental robotics" (Lungarella et al., 2003; Cangelosi and Schlesinger, 2015), which seeks to implement developmental principles in behaving robots to both make robots smarter and test developmental theories "in silicio". The availability of humanoid robots that share even basic body and sensorimotor characteristics with humans opens enormous possibilities to empirically test and improve developmental theorizing that is based on sensorimotor experience. Basic cognitive research on the self has strongly benefited from the establishment of rather simple and easy to implement paradigms, like Botvinick and Cohen's (1998) rubber hand technique, the full-body version (Petkova and Ehrsson, 2008), and the combination of the stroking technique with the visual morphing of faces (Tsakiris, 2008). Additional flexibility was provided by using virtual reality, data gloves, and advanced motion registration (Slater et al., 2010), which allows studying sensorimotor experience under very natural conditions.

The aim of the present Research Topic was to probe the level and the ambitions of current theorizing about the human minimal self. How far did we get? In particular, how far did we get in understanding the mechanistic basis of the human minimal self? How does it emerge? How is it represented? Is it stable or can it be made to disappear, as Buddhist meditation promises? These questions can hardly be answered by pointing to a particular brain area or a particular functional system of which neither the responsible codes nor the operations are specified. What is needed are theoretical assumptions that are sufficiently specific to implement them into an artificial agent and to see whether it can be made to have a self. We were thus interested in any contribution to this question, be it a theoretical comment that synthesizes available research, a review, a particular cognitive, developmental, or other kind of empirical study, a computer simulation, or a robot creating a self. Eleven contributions accepted this challenge and were selected for publication in the Research Topic.

Three Reviews summarize research highlighting the interactive roles of language and interaction, affective processing and agency, and self-other overlap and perspective taking. More specifically, Röder et al. review findings and suggested mechanisms for the grounding of language in the literature on ideomotor theory and identified computational methods that implement decision-making and verbal interaction. They outline how the available computational methods can be used to create advanced computational interaction models that integrate language grounding with body schemas and self-representations.

Kaiser et al. review the available empirical findings on how affective information modulates the experience of agency and how the sense of agency modulates the processing of affective action outcomes. They also discuss whether agency-related changes in affective processing influence the ability to enact cognitive control and action regulation during goal-directed behavior. The authors present a preliminary model that describes the interplay between sense of agency, affective processing, and action regulation. They suggest that affective processing could mediate between subjective sense of agency and the objective ability to control one's behavior.

Müsseler et al. review the available evidence on affective, cognitive, and visuo-spatial perspective taking of humans when facing or working with an avatar. They emphasize that these processes strongly depend on perceived self-avatar overlap or identification with the avatar. They discuss findings showing that when users do take the avatar's perspective, they can show spontaneous behavioral tendencies that run counter to their own.

A Mini Review by Musculus et al. addresses interoception as a crucial aspect of human minimal self in development. Extending on the embodied account of interoceptive inference, the authors present a comparative view of current theoretical frameworks explaining the link between interoception and minimal self. They propose a bi-directional link between motor and interoceptive states that jointly contribute to the formation of minimal self-early on in life. Building upon empirical findings on the development of interoception, they provide an outlook for future research addressing the knowledge gap on interoception in development.

Two Hypothesis and Theory articles address components of the minimal self. Liesner and Kunde focus on the idea of how perceptual changes (e.g., visual, auditory or proprioceptive) that are controllable by efferent activity are considered to be a part of the self. They argue that although this is highly relevant to explaining the experience of agency, sense of body ownership calls for a more nuanced distinction between proprioceptive or tactile (i.e., interoceptive) events and other controllable perceptual events.

Hommel is asking the question how people represent themselves. He proposes that they do so not any differently from how they represent other individuals, events, and objects: by binding codes representing the sensory consequences of being oneself into what he calls a *Me-File*, an event file integrating all the codes resulting from the behaving me. This approach amounts to a Human bundle-self theory of selfhood and uses recent extensions of the Theory of Event Coding (Hommel et al., 2001) for specifying the mechanisms underlying bundle-self-representation.

Two Perspective articles provide further theoretical considerations of how selves might represent themselves. Forch and Hamker discuss how two separate disciplines, namely cognitive science and cognitive robotics, approach the study of minimal self. They argue that whereas cognitive science focus on abstract models predicting and explaining empirical data obtained from humans, cognitive robotics aims at building embodied learning machines that are capable of forming a self similar to humans, which allows researchers to investigate the mechanisms underlying the emergence of the minimal self. They address the differences between human minimal self and robotics models, and provide solutions on how to create models explaining real world behavior.

Bliek et al. extend existing Bayesian models on the embodiment of physically intact limbs to amputated individuals to explain limb embodiment in structurally varying bodies. They focus on the differences in the peripersonal space, limb awareness, the use of prosthetic limbs and sensorimotor learning processes as modulators of the embodiment of artificial limbs in amputated individuals. Combining evidence from neuropsychological research with their modeling approach, they discuss implications of their approach for basic research and clinical contexts.

Three Research articles round up the Research Topic. Adam et al. examine the role of agentive experience and perceptual information on infants' processing of others' action goals. Results show that whereas 7-month-old infants did not show predictive gaze shifts, 18-month-olds did. Moreover, 11-month-olds performed predictive gaze shifts only when a salient action effect was presented. These findings point at a systematic interplay between experience-based top-down processes and cue-based bottom-up information in the development of agentive self-early on in life.

Aerdker et al. report the findings of a developmental psychological study on infant behavior concerning habituation and dishabituation in motor behavior. The study employs the experimental procedure of the habituation paradigm in a movement task to repeated action-effect situations. The experimental results provide evidence for habituation of movement generation that is specific to the direction of the movement, which supports a unified account for patterns of preferential selection based on familiarity preference. Further the authors provide a neural dynamic model that supports experimental results qualitatively and agrees with prior views regarding perceptual habituation.

Finally, Langer and Ay analyze goal-directed action from an information theoretical perspective, by measuring different information flows among the body, the brain, and the environment of an agent. They combine two theories: integrated information theory (related to measures of the amount of information integrated in the controller of the agent in order to quantify consciousness) and morphological computation (which refers to the problem from an exterior viewpoint by analyzing how the morphology of the agent and its interaction with the environment can lift the computational burden of the brain). In their experimental case study, they observe an antagonistic relationship between morphological computation and integrated information.

Overall, the Research Topic brings together different positions that contribute to current theorizing about the human minimal self. It paves the way for interdisciplinary work and will stimulate further research on how people represent themselves.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Botvinick, M., and Cohen, J. (1998). Rubber hands feel touch that eyes see. *Nature.* 391, 756–756. doi: 10.1038/35784

Cangelosi, A., and Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. MIT Press.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Gredebäck, G., Johnson, S., and von Hofsten. (2010). Eye tracking in infancy research. *Dev. Neuropsychol.* 35, 1–19. doi: 10.1080/87565640903325758

Hafner, V. V., Loviken, P., Pico Villalpando, A., and Schillaci, G. (2020). Prerequisites for an artificial self. *Front. Neurorobotics.* 14:5. doi: 10.3389/fnbot.2020.00005

Hommel, B. (2016). "Embodied cognition according to TEC," in *Foundations of Embodied Cognition*, Volume 1: Perceptual and Emotional Embodiment, eds Y. Coello and M. Fischer (Psychology Press), 75–92.

Hommel, B. (2017). "Goal-directed actions," in *Handbook of Causal Reasoning*, ed M. Waldmann (Oxford: Oxford University Press). doi: 10.1093/oxfordhb/9780199399550.013.18

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–878. doi: 10.1017/S0140525X01000103

Kilner, J., Hommel, B., Bar, M., Barsalou, L. W., Friston, K. J., Jost, J., et al. (2015). "Action-oriented models of cognitive processing: A little less cogitation, a little more action please," in *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science*, eds A. K. Engel, K. J. Friston, and D. Kragic (Cambridge, MA: MIT Press), 159–172.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *ConSci.* 15, 151–190. doi: 10.1080/09540090310001655110

Paulus, M. (2022). Should infant psychology rely on the violation-of-expectation method? *Not anymore. Infant Child Dev.* 31, e2306. doi: 10.1002/icd.2306

Petkova, V. I., and Ehrsson, H. H. (2008). If i were you: perceptual illusion of body swapping. *PLoS ONE.* 3, e3832. doi: 10.1371/journal.pone.0003832

Saby, J. N., and Marshall, P. J. (2012). The utility of EEG band power analysis in the study of infancy and early childhood. *Dev. Neuropsychol.* 37, 253–273. doi: 10.1080/87565641.2011.614663

Schillaci, G., Hafner, V. V., and Lara, B. (2016a). Exploratiohaviours, body representations and simulation processes for the development of cognition in artificial agents *.Front. Robotics AI: Humanoid Robotics.* 3, 39. doi: 10.3389/frobt.2016.00039

Schillaci, G., Ritter, C.-N., Hafner, V. V., and Lara, B. (2016b). Body representations for robot ego-noise moing and prediction. *Towards the Development of a Sense* *of Agency in Artificial Agents, International Conference on the Simulation and Synthesis of Living Systems* (ALife XV). Mexico. p. 390-397.

Slater, M., Spanlang, B., Saz-Vives, M. V., and Blanke, O. (2010). First person experience of body transfer in virtual reality. *PLoS ONE.* 5:e10564. doi: 10.1371/journal.pone.0010564

Tsakiris, M. (2008). Looking for myself: Current multisensory input alters self-face recognition. *PLoS ONE* 3, e4040. doi: 10.1371/journal.pone.0004040

Vernon, D., Beetz, M., and Sandini, G., (2015). Prospection in cognition: the case for joint episodic-procedural memory in cognitive robotics. *Front. Rocs and AI: Humanoid.* 2, 19 doi: 10.3389/frobt.2015.00019

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

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

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 30 July 2021
doi: 10.3389/fpsyg.2021.698778

The Me-File: An Event-Coding Approach to Self-Representation

Bernhard Hommel[1,2,3]*

[1] *Cognitive Psychology Unit, Institute for Psychological Research & Leiden Institute for Brain and Cognition, Leiden University, Leiden, Netherlands,* [2] *Cognitive Neurophysiology, Department of Child and Adolescent Psychiatry, Faculty of Medicine, TU Dresden, Dresden, Germany,* [3] *Department of Psychology, Shandong Normal University, Jinan, China*

Numerous authors have taken it for granted that people represent themselves or even have something like "a self", but the underlying mechanisms remain a mystery. How do people represent themselves? Here I propose that they do so not any differently from how they represent other individuals, events, and objects: by binding codes representing the sensory consequences of being oneself into a Me-File, that is, into an event file integrating all the codes resulting from the behaving me. This amounts to a Humean bundle-self theory of selfhood, and I will explain how recent extensions of the Theory of Event Coding, a general theory of human perception and action control, provide all the necessary ingredients for specifying the mechanisms underlying such a theory. The Me-File concept is likely to provide a useful mechanistic basis for more specific and more theoretically productive experimentation, as well as for the construction of artificial agents with human-like selves.

Keywords: self representation, agency, body ownership, Theory of Event Coding (TEC), minimal self

OPEN ACCESS

Edited by:
Jutta Kray,
Saarland University, Germany

Reviewed by:
Herbert Heuer,
Leibniz Research Centre for Working Environment and Human Factors (IfADo), Germany
Julia Karbach,
University of Koblenz and Landau, Germany

Correspondence:
Bernhard Hommel
bh@bhommel.onmicrosoft.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *22 April 2021*
Accepted: *30 June 2021*
Published: *30 July 2021*

Citation:
Hommel B (2021) The Me-File: An Event-Coding Approach to Self-Representation. Front. Psychol. 12:698778. doi: 10.3389/fpsyg.2021.698778

INTRODUCTION

Like many other concepts used in academic psychology, the concept of the "self" is rather uncritically taken to refer to something residing in the human mind or brain or both that creates some degree of unity of either the phenomenal experience that we have with or about us or the stories that we are telling about us. Nowhere does one find the concept to be questioned or justified, apparently because both authors and readers consider the existence of a self self-evident (pun partly intended). The reason for this uncritical acceptance is likely to be its philosophical heritage: the only toolbox that philosophers traditionally have available to acquire their data is themselves and their phenomenal experience, so that it does not come as a surprise that the only thing that Descartes was unable to doubt was (the phenomenal experience of) the doubting self. However, less subjective methods did not provide strong support for our intuition that our phenomenal experience plays an important role in our thinking and acting, as it turned out to be too slow and too error-prone to represent a promising causal factor in human perception and action (Nisbett and Wilson, 1977; Wegner, 2002; Hommel, 2013). Moreover, the mere fact that a concept exists in our language cannot be taken as existence proof for a dedicated psychological mechanism responsible for generating the behavior this concept refers to (Danziger, 1997). More specifically, while there is nothing wrong with categorizing all information that receptors provide about the agent carrying them as "belonging to or constituting a self," the mere fact that this information can be consciously perceived does not yet require any mechanism creating any unity. Along the same lines, the fact that people tend to play the main role in their narratives does not require any dedicated mechanism that makes sure that they do—it may simply be the fact that they happen to be the one they are the most familiar with.

These considerations raise the suspicion that the self-concept carries quite a bit of unnecessary baggage that reflects the natural bias that a limitation of one's empirical toolbox to self-experience brings with it, rather than straightforward functional considerations calling for a dedicated self-mechanism. They also raise the suspicion that many theorists are not yet decided whether they consider the self in its various disguises an explanandum that their theory aims to explain or an explanans that provides this explanation. In fact, many theories try to explain the self as explanandum by referring to some not further explained internal self-system that has no other purpose than generating the explanandum—a clear case of pseudo-explanation (Hommel, 2020). In the following, my aim will be to drop this baggage and develop a purely functional theoretical approach to what we call the self. That is, my aim will be to explain the behavior that theorists consider reflections of a self without referring to a dedicated system producing that behavior. In fact, I will try to do without inventing any new mechanisms to account for such behavior and restrict myself to the Theory of Event Coding (TEC; Hommel et al., 2001; Hommel, 2019a) as my theoretical toolbox.

TEC was conceived as a generic theory of the representations and processes underlying human perception and action. It assumes that perceived and produced events (i.e., action plans) are represented by bindings of codes representing the features of these events, so-called event files (Hommel, 2004). First versions addressed perception and action in very simple tasks involving stimuli with very few features, like red circles and green rectangles, and not overly complex actions, like pressing left and right keys. However, more recent versions addressed more complex tasks and situations (Hommel, 2019a) and questions of self- and other-representation (Hommel, 2018) by means of the same mechanistic principles. Indeed, the representational assumptions of TEC are fully consistent with theoretical frameworks targeting more social processes, including self-representation (Greenwald et al., 2002), which is why I consider the mechanistic toolbox of TEC fully sufficient for understanding self-representation, despite the theory's non-social origin.

ONLINE AND OFFLINE SELF

Psychological approaches to the self commonly accept the philosophical distinction between minimal and narrative self. And indeed, it makes intuitive sense to distinguish between Hume's 1739 idea of a personal self consisting of nothing but the perceptual information that an agent has available about herself, so that she in some sense "ceases to exist" when falling asleep, and the idea of an agent who actively sculpts the image of herself by telling self-relevant stories (Gergen and Gergen, 1997; Gallagher, 2000). However, this distinction is heavily confounded with various other factors: the timeframe (second by second versus minutes or years), the medium (perception versus communication), the audience (oneself versus oneself vis-à-vis others), and the reliance on earlier experience, so that it remains unclear whether the distinction between minimal and narrative self actually refers to different concepts, different mechanisms, different kinds of experience, or something else. From a purely functional viewpoint, it seems more reasonable, so I suggest, to distinguish between online and offline self.

Online Self

The *online self* refers to the here and now, to the flow of information from receptors to more integrative processing levels that inform action control, and vice versa. According to TEC, a person would represent herself just like any other event: by a binding of codes representing the features making up the event, oneself in this particular case. This comprises of all perceivable features regarding oneself in principle, features referring to how one looks, sounds, and smells, but also how one moves and feels—which reflects the ideomotor heritage of TEC, according to which actions and emotions are also grounded in self-perception. Which features belong to this "personal" event may not always be obvious. For instance, infants need quite a while before they develop a good understanding of which objects and events do or do not belong to themselves, and active exploration of their own body and their immediate surrounding plays an important role in this development (for a review, see Verschoor and Hommel, 2017). Even adults can be surprisingly flexible in their self-perception, as indicated by the notorious rubber-hand illusion (Botvinick and Cohen, 1998): when participants are confronted with a rubber hand lying in front of them, simultaneously stroking the rubber hand and the participant's real hand results in the illusion that the rubber hand becomes part of the participant's own body.

These observations suggest that people are not born with a fixed representation of themselves but continuously re-create their self-representation based on the currently available perceptual information. To determine whether perceived features are actually related to themselves or to their physical or social environment, people seem to use the same cues that are known from object perception. For instance, people are more likely to perceive rubber or virtual hands as part of their own body if these artificially effectors are spatially close to their body, if they can be seen as a continuation of their own effectors, and if artificial and real effectors move in synchrony (e.g., Ma and Hommel, 2015). In object and non-social event perception, these kinds of cues are known as the Gestalt laws of spatial and temporal proximity, good Gestalt/continuation, and common fate (Todorovic, 2008), which supports the idea that representing oneself follows the same principles as representing other events. Another well-known principle governing self-perception is the relationship between intended and actual action effects (Hommel, 2015): the event with the closest relationship (i.e., the one that keeps generating action effects that I intend) is probably me (Verschoor and Hommel, 2017). This relationship is an important ingredient of any control system, ranging from central heating to human intentional action (Frith et al., 2000), and presumably the crucial information for judging personal agency (Blakemore et al., 2002).

While the online self can be informed by and interact with stored information (the activated bits of the offline self), it is mainly a reflection of the incoming, currently available information that active agents generate themselves. Accordingly, the binding of the codes that represent the features that specify the active agent—the structure that I will call the Me-file—can be considered to represent the self as envisioned by Hume's bundle-theory, that is, as a direct perceptual reflection of how we currently embody ourselves. Note that this reflection does not distinguish between cognitive, motivational, and affective (or any other kind of) information. As elaborated elsewhere (Hommel, 2019b), such labels refer to different functions of representations and mechanisms but do not necessarily indicate that the underlying representations and mechanisms themselves are separable and specific. For instance, Barrett (2017) has argued that perceived emotion and affect are not generated by dedicated affective mechanisms but derived from general mechanisms with basic survival functions, so that it makes little sense to consider the mechanisms as cognitive, motivational, or affective. Along these lines, the online Me-file of a jogging colleague might look like in the left panel of **Figure 1**, where going for a jog provides her with feedback about her being busy with running, with being athletic, with being short and female, but also with being happy—among many other features that online feedback might inform about.

Offline Self

If Hume is right in claiming that people in some sense cease to exist when going asleep, and if this scenario is taken to reflect the fact that we more or less switch off online self-perception during the night, it is easy to see that the online self cannot be all that we have. Obviously, people do not start from scratch in perceiving themselves when waking up, which means that we are able to store perceived information about ourselves in a more durable format—the offline self. As suggested by self-perception theorists like Bem (1972) and Laird (2007), people learn about themselves just like they learn about others: by perceiving their behavior and looking for regularities. I thus do not have privileged information about me being a friendly or aggressive person, say, but I may assume being one if I perceive myself to repeatedly compliment other people or punch them in the face, respectively. Repeatedly making such observations and representing them in my online self is likely to leave traces behind, traces that survive the switching off of my online self during sleep and that provide me with a warm-start the next morning. Accordingly, Greenwald et al. (2002) have suggested that people keep networks of feature codes that refer to one's perceived personal characteristics, like being athletic, intelligent, a professor, grandmother, female, and short, as indicated in the right panel of **Figure 1**. In contrast to the online self, which is restricted to those feature codes that are currently activated (for reasons discussed in the next section), the offline self refers to the total of all available feature codes that have been involved in self-representation to a degree that they have been bound into a network that represent something like the potential self. In other words, the offline self refers to the knowledge that a person has acquired about herself, about the features that she knows to have in principle.

Current Self

It is important to emphasize that the terms online self and the offline self do not imply different systems but refer to different levels of activation of feature codes. Cowan (1995) has suggested that short-term memory might be considered the activated part of long-term memory. Hence, whereas long-term memory contains all codes that a person has acquired over the years, only some of these codes are active at any time, irrespective of whether they have been exogenously or endogenously activated, and the total of the currently active codes constitute short-term memory. The same applies to self-representation. The offline self is the total of all feature codes that have become part of the network of codes that the person has learned to represent features of her and that she has used to represent herself in the past. According to TEC, being exposed to a situation, being engaged in a task, and being busy with particular themes increases the *intentional weighting* (Hommel et al., 2001; Memelink and Hommel, 2013) of feature dimensions that the agent considers relevant (based on past experience and current expectations) for making the right choices under these situational circumstances. This means that feature values that are coded on these dimensions are activated more strongly and have a higher impact to impact decision-making and action-selection. If, thus, a participant is asked to press a left versus right key in response to red and green stimuli, respectively, the intentional weighting for color and location will be high, given that these dimensions define the task-relevant aspects of stimuli and responses. Indeed, even preparing for simple tasks like pointing, grasping, and tapping is sufficient to sensitize the agent for attending to and prioritizing stimuli falling onto dimensions that are important for these actions, like location, shape, and rhythm (Schubotz and von Cramon, 2003; Fagioli et al., 2007).

With respect to self-representation, this means that the way we currently represent ourselves is selective and strongly affected by our current concerns and interests, the tasks we carry out, and the situational implications they have. Our *current self* would thus be a mixture of codes that represent currently perceived features of ourselves (the online self), in particular of features related to dimensions that we currently consider relevant, and those feature codes of our offline self that are active for other reasons, perhaps because they are relevant for another task we pursue or intend to pursue in the near future, or because of our current concerns (Klinger and Cox, 2011)—thoughts we are busy with, or because of other needs, like hunger or a need for affiliation (McClelland, 1988; Hommel, 2021). This implies that we are not always the same and do not perceive ourselves as the same under all circumstances. Entering particular social bubbles, like when visiting or family or meeting old friends, is likely to implement different sets of intentional weighting, which in turn will emphasize particular features in our self-perception and deemphasize others. With respect to our example described above, participating in a running competition would increase the intentional weighting for features related to being sporty and fast, so that self-perception would focus on information that is likely to activate the feature codes for being athletic and running, but probably not feature codes for being a professor or a wife.

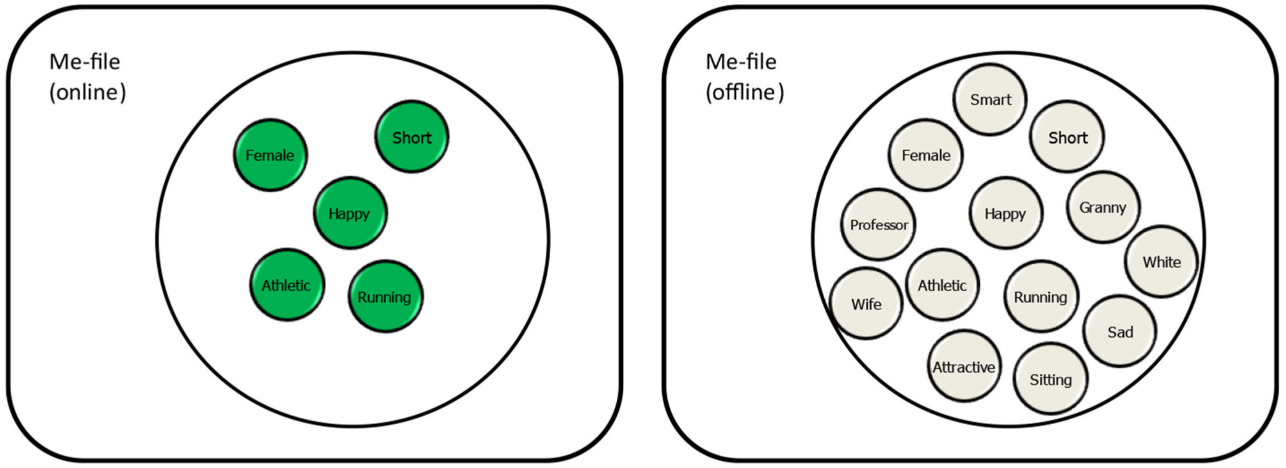


FIGURE 1 | Online and offline Me-files including currently activated (in green, see left panel) and currently inactive (in gray, see right panel) feature codes characterizing the agent.

THEORETICAL IMPLICATIONS

In essence, my claim is that people represent themselves like any other event, so that no special theoretical claims need to be made, no novel mechanisms need to be introduced, and no additional assumptions need to be defended, to account for our ability to represent ourselves. And yet, my minimalist account has interesting theoretical implications that can account for numerous empirical observations that have either not been sufficiently well explained so far or that have been explained with specialized, and thus not overly parsimonious theoretical frameworks. In the following, I will briefly touch some of these implications and phenomena they relate to.

Ownership

My account does not assume any dedicated mechanism responsible for perceiving body ownership, as when being confronted with a body extension, be it a tool or an artificial hand. Instead, it assumes that people judge the degree to which an artificial hand belongs to their body in exactly the same way as they judge the relationship between a dog and its tail: if the tail is close to the dog, if it wiggles only when the dog moves as well, and if it tends to appear and disappear together with the dog, people will perceive the tail as part of the dog. The same applies to a rubber or virtual hand: if it is close to me, if it moves when I move, and if it accompanies me wherever I go, I'm likely to consider it as part of me and my body. Obviously, the informational basis for judging the relationship between oneself and a candidate body part is different from judging the relationship between someone else and a candidate body part: visual information tends to be more comprehensive when observing other agents, whereas interoceptive (kinesthetic, proprioceptive) and tactile information will commonly be available only when perceiving oneself. This may mean that the outcomes of such judgments rely on different kinds of information and may be difficult to compare. Nevertheless, this does not imply any difference in the way the available information is integrated and analyzed, which means that the basic mechanisms and their principles do not differ. It is certainly true that this account does not yet address all theoretical questions. Most importantly, why is it these Gestalt criteria (spatial/temporal proximity, good Gestalt/continuity, and common fate) that people tend to use when judging relationships between events? Are these simply the most reliable indicators or are there cultural or educational factors involved? Tackling such questions is an important challenge for future research, but it is not a question that would be specific for self-representation.

Agency

Judgments of body ownership and agency tend to be dissociable in the highly artificial rubber-hand scenarios but are strongly correlated in studies with more natural relationships between real body movements and movements of artificial extensions (Ma et al., 2019). This suggests that the informational basis for judging agency and judging body ownership overlaps to a substantial degree. However, there is substantial evidence for a special role of the relationship between personal intentions and related expectations of action outcomes on the one hand and the actual outcomes on the other for judging agency (Hommel, 2015). There is theoretical consensus that information about this relationship can be directly derived from mechanisms underlying action control. Voluntary action is assumed to be selected based on expected action outcomes, which is almost true by definition: given that voluntary action is defined as aiming at particular outcomes, representations of outcomes must play some role in selecting the movements that eventually achieve these outcomes (Hommel, 2009). Moreover, adaptive action control requires insight into the degree to which a particular action has or has not generated the intended action effects, and this insight is commonly derived from comparing expected outcomes with actual outcomes (Frith et al., 2000). It is the result of this comparison that is assumed to contribute to judgments of agency

(Blakemore et al., 2002; Chambon and Haggard, 2013; Hommel, 2015), which again means that accounting for agency does not need any dedicated mechanism beyond what has to be assumed for voluntary action control anyway.

Sticky Intentions

Various authors have pointed out that committing oneself to a goal or intention makes it particularly sticky (Hollenbeck and Klein, 1987). Lewin (1936) suggested that committing oneself to a goal creates a kind of tension in one's cognitive system that seeks for relaxation very much like a biological drive seeks for reduction. Along the same lines, Klinger (2013) suggests that self-commitment turns mere motivation into goal-striving which, among other things, keeps the respective goal active until the intended outcome has been achieved. Commitment to the goal was also considered crucial to engage in actual goal-striving by Locke and colleagues (e.g., Locke et al., 1988) or Gollwitzer and Oettingen (2011), and there is indeed massive evidence suggesting that self-reported commitment to the goal is the central predictor of successful performance, especially in difficult tasks (Hollenbeck and Klein, 1987; Klein et al., 1999). Along the same lines, Goschke and Kuhl (1993) and others demonstrated that concepts that are connected to actual goals are much easier to remember than concepts that are not (intention memory). The authors suggested that this might be due to some special kind of energy that keeps goal-related representations more active than others—but what this special energy (or Lewin's cognitive tension) might consist of remains a mystery.

From a Me-file perspective, the consideration of two well-established mechanistic features of our cognitive system is sufficient to account for sticky intentions. First, preparing for a task allows people to create lasting associations between task-relevant representations. Hence, if, for instance, participants are instructed to carry out action X in response to stimulus A and action Y in response to stimulus B, they seem to create bindings between the representations of A and X and between the representations B and Y even before the very first trial, as witnessed by the observation that, after the instruction has been given, stimuli acquire the power to automatically activate the response they have been assigned to (Meiran et al., 2017). Second, given that every movement of ours provides perceptual feedback about us, our online self is always active, at least as long as we are awake, and so is our current self of which the online self is a part. If so, each feature code that is part of the current self must also be consistently primed to at least some degree, depending on the degree of intentional weighting. Connecting these two considerations suggests that the act that phenomenologically consists in committing to a goal or intention reflects the mechanistic process of merging the representation of this goal/intention with the Me-file (similar to the assumption of Salancik, 1977, that commitment represents a kind of binding between an individual and her actions). As elaborated elsewhere, goals are likely to be represented by criteria that constrain the selection of event files in such a way that goal-consistent actions become more likely to be selected (Hommel and Wiers, 2017; Hommel, 2021). Accordingly, committing to a goal would integrate corresponding selection criteria into the Me-file. As the Me-file tends to be active most of the time, so would the goal criteria, which would explain why not yet achieved goals are sticky—without referring to any metaphorical tension or mysterious energy.

Self-Symbols

The consideration that associating information with the Me-file could make that information more accessible and increase its impact on selection might also account for a not yet fully understood observation of Sui, Humphreys, and colleagues (e.g., Sui et al., 2012; Sui and Humphreys, 2015). These authors presented participants with arbitrary symbols and asked them to associate these symbols with either themselves, a close relative, or a stranger, before presenting the symbols in simple cognitive tasks. It turned out that the self-related symbol was responded to faster and recall better in various kinds of tasks, suggesting that the simple fact that a symbol was taken to refer to the participant was sufficient to make that symbol enjoy highly prioritized processing. Considering that the instruction to associate a symbol with oneself might consist in integrating that symbol into one's more or less consistently active Me-file would easily account for the reported observations.

Resting State

The idea of a chronically active Me-file would also fit with the observation that cortical midline regions involved in *resting-state* or *default-mode* activity (i.e., the typical neural activity shown in the absence of a particular task) show strong spatial overlap with regions that are recruited during self-referential processing (D'Argembeau et al., 2005; Qin and Northoff, 2011). The typical instruction in resting-state studies asks participants to engage in no particular task or thought. To the degree that participants follow this instruction, all that remains will be sensory feedback about themselves, which in turn will activate codes that are contained in the Me-file and contribute to the chronically high level of activation of that file. If so, it is easy to understand why this activates areas that are also active during intentional self-referential processing.

Social Discrimination

Recent political discussions often focus on aspects of social discrimination, be they related to the proper representation or treatment of people with a particular gender, skin color, political or religious orientation, or sexual preference. There are basically two ideas of how discrimination related to any of these features might be overcome: by reducing/eliminating possible or actual attention to the underlying feature dimension (e.g., as implied by the so-called color-blindness theory: Ansell, 2013) or by increasing attention to this dimension (e.g., as claimed by the Woke movement: en.wikipedia.org/wiki/Woke). It might be interesting to mention that my approach suggests concrete hypotheses regarding the processes that these two strategies would evoke and which consequences they would have. Having the goal of attending to skin color would be likely to create a strong association between the codes representing that feature and one's Me-file. This would render skin color an important feature to represent oneself and others, and be likely to make skin color a feature dimension that overshadows other possible dimensions, like those coding for gender, achievement, sociality, and more. Given that discrimination can be positive or negative, depending on one's experience and values, this does not allow predicting the exact consequences. But my approach would predict that Woke principles should increase and stabilize both the absolute and the relative (as compared to other feature dimensions) importance of the targeted feature dimension in perception (of oneself and others), decision-making, and action—which provides a continuous basis for discriminative behavior.

Individual Differences

The Me-file approach to self-representation provides a novel perspective on inter- and intra-individual differences in self-perception and the impact of self-perception on behavior (or vice versa). As discussed in the previous section, different physical and social contexts are likely to moderate the intentional weighting of both perceptual dimensions and particular context-specific themes. For instance, going to the gym or participating in a sports event in a sense "reduces" the self-perceiving individual to her physical, performance-relevant attributes and abilities, downplaying other aspects, like gender, race, wealth, and academic background, whereas visiting a library will highlight very different attributes and abilities. Spending time with one's peer groups will increase the weight of other perceptual dimensions and themes than spending time with one's parents, which in turn is not unlikely to change one's behavior and the way one perceives oneself. One of the many interesting aspects of these considerations refers to retirement. As discussed by Hommel and Kibele (2016), an important aspect of cognitive aging (i.e., the decline of cognitive abilities with increasing age) is likely to do with what might be called the embodiment of (non-)agency: Retirement is commonly accompanied by a sudden and rather extensive reduction of one's action repertoire and of the opportunities to experience oneself as being an agent that makes active use of this repertoire. The Me-file approach suggests that this must lead to a drastic reduction of the complexity of self-representation, as the individual no longer perceives herself as an active agent in the physical and social world in quite a number of situations—the kind and number of which depends on the particular job one retires from. Hence, not only is the retired individual prevented from actively exercising the cognitive skills the previous job required, but she is also unlearning to perceive herself as someone who does these things: a kind of acquired non-agency. If so, forced retirement might be considered a societal act that undermines personal motivation and self-respect. Other implications refer to upbringing and education. If, as the Me-file suggests, action is such an important ingredient of self-representation, explorative, active learning would not only be mandated for possible educational reasons but also for the building of active self's, that is, for identities that include the agentive aspect of individuals.

CONCLUSION

My aim was to present a mechanistically transparent basis for theorizing about the human self. I have used TEC as my theoretical toolbox and argued that no dedicated special assumptions or principles need to be added to account for self-representation. More specifically, I suggest that representing oneself follows the exact same principles as representing others or representing things, even though the type and the amount of information that is available for the resulting representations is likely to differ—for obvious and theoretically not overly relevant reasons, like the fact that some sensory channels provide more information about oneself than about others. I have also suggested that what philosophical approaches have considered the key ingredients of the human self—body ownership and agency—do not require any special theorizing or any dedicated system or mechanism. In fact, reports about body ownership and agency are likely to be based on the same principles that underlie the judgment of relatedness and causality regarding non-personal events, like the motions of billiard balls and, in the case of agency, on comparisons between intended and actual action effects, as available from action-control processes. Hence, what we call the self may not be special at all, and not require any special theorizing. Given that humans are both subjects and objects of research on the self, this may be intellectually disappointing, especially when viewing the issue from the object perspective. However, it does allow us to create mechanistically transparent models that do not require any special modules or systems to account for the selfness aspect of representing ourselves. In particular, the approach allows implementing various aspects of human-like selfhood into various kinds of artificial agents, and even constructing agents that spontaneously acquire their self through sensorimotor experience with their own embodiment.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was supported by a grant from the German Research Community (DFG; HO 1430/12-1) and a 100 Talent Grant of the Province of Shandong, China.

REFERENCES

Ansell, A. E. (2013). Race and ethnicity: the key concepts. New York: Routledge. doi: 10.4324/9780203448236

Barrett, L. F. (2017). *How Emotions are Made: The Secret Life of the Brain*. New York: Houghton Mifflin Harcourt.

Bem, D. J. (1972). Self-perception theory. *Adv. Exp. Soc. Psychol.* 6, 1–62. doi: 10.1016/S0065-2601(08)60024-6

Blakemore, S. J., Wolpert, D. M., and Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends Cogn. Sci.* 6, 237–242. doi: 10.1016/S1364-6613(02)01907-1

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature* 391, 756–756. doi: 10.1038/35784

Chambon, V., and Haggard, P. (2013). "Premotor or ideomotor: how does the experience of action come about?," in *Action Science: Foundations of an Emerging Discipline,* eds W. Prinz, M. Beisert and A. Herwig (Cambridge, MA: MIT Press), 359–380. doi: 10.7551/mitpress/9780262018555.003.0014

Cowan, N. (1995). *Attention and Memory: An Integrated Framework*. New York, NY: Oxford University Press.

Danziger, K. (1997). *Naming the Mind: How Psychology Found Its Language*. Thousand Oaks, CA: Sage Publications.

D'Argembeau, A., Collette, F., Van der Linden, M., Laureys, S., DelFiore, G., Degueldre, C., et al. (2005) Self-referential reflective activity and its relationship with rest: a PET study. *Neuroimage* 25, 616–624. doi: 10.1016/j.neuroimage.2004.11.048

Fagioli, S., Hommel, B., and Schubotz, R. I. (2007). Intentional control of attention: action planning primes action-related stimulus dimensions. *Psychol. Res.* 71, 22–29. doi: 10.1007/s00426-005-0033-3

Frith, C. D., Blakemore, S. J., and Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 355, 1771–1788. doi: 10.1098/rstb.2000.0734

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Gergen, K. J., and Gergen, M. M. (1997). "Narratives of the Self," in *Suny Series in the Philosophy of the Social Sciences. Memory, Identity, Community: The Idea of Narrative in the Human Sciences* (New York, NY: State University of New York Press), 161–184.

Gollwitzer, P. M., and Oettingen, G. (2011). "Planning promotes goal striving," in *Handbook of Self-Regulation: Research, Theory, and Applications, 2nd ed*, eds K. D. Vohs, and R. F. Baumeister (New York: Guilford), 162–185.

Goschke, T., and Kuhl, J. (1993). Representation of intentions: persisting activation in memory. *J. Exp. Psychol. Learn. Memory Cogn.* 19, 1211–1226. doi: 10.1037/0278-7393.19.5.1211

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., and Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychol. Rev.* 109, 3–25. doi: 10.1037/0033-295X.109.1.3

Hollenbeck, J. R., and Klein, H. J. (1987). Goal commitment and the goal-setting process: problems, prospects, and proposals for future research. *J. Appl. Psychol.* 72, 212–220. doi: 10.1037/0021-9010.72.2.212

Hommel, B. (2004). Event files: feature binding in and across perception and action. *Trends Cogn. Sci.* 8, 494–500. doi: 10.1016/j.tics.2004.08.007

Hommel, B. (2009). Action control according to TEC (theory of event coding). *Psychol. Res.* 73, 512–526. doi: 10.1007/s00426-009-0234-2

Hommel, B. (2013). Dancing in the dark: no role for consciousness in action control. *Front. Psychol.* 4:380. doi: 10.3389/fpsyg.2013.00380

Hommel, B. (2015). "Action control and the sense of agency," in *The Sense of Agency*, eds P. Haggard, and B. Eitam (New York: Oxford University Press), 307–326. doi: 10.1093/acprof:oso/9780190267278.003.0014

Hommel, B. (2018). Representing oneself and others: an event-coding approach. *Exp. Psychol.* 65, 323–331. doi: 10.1027/1618-3169/a000433

Hommel, B. (2019a). Theory of event coding (TEC) V2.0: representing and controlling perception and action. *Atten. Percept. Psychophys.* 81, 2139–2154. doi: 10.3758/s13414-019-01779-4

Hommel, B. (2019b). Affect and control: a conceptual clarification. *Int. J. Psychophysiol.* 144, 1–6. doi: 10.1016/j.ijpsycho.2019.07.006

Hommel, B. (2020). Pseudo-mechanistic explanations in psychology and cognitive neuroscience. *Top. Cogn. Sci.* 12, 1294–1305. doi: 10.1111/tops.12448

Hommel, B. (2021). *GOALIATH: A Theory of Goal-Directed Behavior* (Submitted).

Hommel, B., and Kibele, A. (2016). Down with retirement: implications of embodied cognition for healthy aging. *Front. Psychol.* 7:1184. doi: 10.3389/fpsyg.2016.01184

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–878. doi: 10.1017/S0140525X01000103

Hommel, B., and Wiers, R. W. (2017). Towards a unitary approach to human action control. *Trends Cogn. Sci.* 21, 940–949. doi: 10.1016/j.tics.2017.09.009

Hume, D. (1739). *A Treatise of Human Nature*. Available online at: https://librivox.org/treatise-of-human-nature-vol-1-by-david-hume (accessed on July 10, 2021).

Klein, H. J., Wesson, M. J., Hollenbeck, J. R., and Alge, B. J. (1999). Goal commitment and the goal-setting process: conceptual clarification and empirical synthesis. *J. Appl. Psychol.* 84, 885–896. doi: 10.1037/0021-9010.84.6.885

Klinger, E. (2013). Goal commitments and the content of thoughts and dreams: basic principles. *Front. Psychol.* 10:415. doi: 10.3389/fpsyg.2013.00415

Klinger, E., and Cox, W. M. (2011). "Motivation and the goal theory of current concerns," in *Handbook of Motivational Counseling, 2nd Ed*, eds W. M. Cox, and E. Klinger (Chichester: Wiley), 3–47. doi: 10.1002/9780470979952.ch1

Laird, J. D. (2007). *Feelings: The Perception of Self*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195098891.001.0001

Lewin, K. (1936). *Principles of Topological Psychology*. New York, NY: McGraw-Hill. doi: 10.1037/10019-000

Locke, E. A., Latham, G. P., and Erez, M. (1988). The determinants of goal commitment. *Acad. Manag. Rev.* 13, 23–39. doi: 10.5465/amr.1988.4306771

Ma, K., and Hommel, B. (2015). Body-ownership for actively operated non-corporeal objects. *Conscious. Cogn.* 36, 75–86. doi: 10.1016/j.concog.2015.06.003

Ma, K., Hommel, B., and Cheng, H. (2019). The roles of consistency and exclusivity in perceiving body ownership and agency. *Psychol. Res.* 83, 175–184. doi: 10.1007/s00426-018-0978-7

McClelland, D. C. (1988). *Human Motivation*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139878289

Meiran, N., Liefooghe, B., and De Houwer, J. (2017). Powerful instructions: automaticity without practice. *Curr. Dir. Psychol. Sci.* 26, 509–514. doi: 10.1177/0963721417711638

Memelink, J., and Hommel, B. (2013). Intentional weighting: a basic principle in cognitive control. *Psychol. Res.* 77, 249–259. doi: 10.1007/s00426-012-0435-y

Nisbett, R., and Wilson, T. (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

Qin, P., and Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *Neuroimage* 57, 1221–1233. doi: 10.1016/j.neuroimage.2011.05.028

Salancik, G. R. (1977). Commitment is too easy! *Organ. Dyn.* 6, 62–80. doi: 10.1016/0090-2616(77)90035-3

Schubotz, R. I., and von Cramon, D. Y. (2003). Functional-anatomical concepts of human premotor cortex: evidence from fMRI and PET studies. *Neuroimage* 20, S120–S131. doi: 10.1016/j.neuroimage.2003.09.014

Sui, J., He, X., and Humphreys, G. W. (2012). Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. *J. Exp. Psychol. Human Percept. Perform.* 38, 1105–1117. doi: 10.1037/a0029792

Sui, J., and Humphreys, G. W. (2015). More of me! Distinguishing self and reward bias using redundancy gains. *Atten. Percept.*

Psychophys. 77, 2549–2561. doi: 10.3758/s13414-015-0970-x

Todorovic, D. (2008). Gestalt principles. *Scholarpedia* 3:5345. doi: 10.4249/scholarpedia.5345

Verschoor, S. A., and Hommel, B. (2017). Self-by-doing: the role of action for self-acquisition. *Soc. Cogn.* 35, 127–145 doi: 10.1521/soco.2017.35.2.127

Wegner, D. M. (2002). *The Illusion of Conscious Will.* Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3650.001.0001

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

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

The Impact of Action Effects on Infants' Predictive Gaze Shifts for a Non-Human Grasping Action at 7, 11, and 18 Months

Maurits Adam[1]*, *Christian Gumbsch*[2,3], *Martin V. Butz*[2] *and Birgit Elsner*[1]

[1]*Developmental Psychology, Department of Psychology, University of Potsdam, Potsdam, Germany,* [2]*Neuro-Cognitive Modeling, Department of Computer Science and Department of Psychology, University of Tübingen, Tübingen, Germany,* [3]*Autonomous Learning Group, Max Planck Institute for Intelligent Systems, Stuttgart, Germany*

During the observation of goal-directed actions, infants usually predict the goal at an earlier age when the agent is familiar (e.g., human hand) compared to unfamiliar (e.g., mechanical claw). These findings implicate a crucial role of the developing agentive self for infants' processing of others' action goals. Recent theoretical accounts suggest that predictive gaze behavior relies on an interplay between infants' agentive experience (top-down processes) and perceptual information about the agent and the action-event (bottom-up information; e.g., agency cues). The present study examined 7-, 11-, and 18-month-old infants' predictive gaze behavior for a grasping action performed by an unfamiliar tool, depending on infants' age-related action knowledge about tool-use and the display of the agency cue of producing a salient action effect. The results are in line with the notion of a systematic interplay between experience-based top-down processes and cue-based bottom-up information: Regardless of the salient action effect, predictive gaze shifts did not occur in the 7-month-olds (least experienced age group), but did occur in the 18-month-olds (most experienced age group). In the 11-month-olds, however, predictive gaze shifts occurred only when a salient action effect was presented. This sheds new light on how the developing agentive self, in interplay with available agency cues, supports infants' action-goal prediction also for observed tool-use actions.

Keywords: infancy, predictive gaze behavior, eye tracking, tool-use actions, agency cues, developing agentive self, non-human grasping

OPEN ACCESS

Edited by:
Stephan Alexander Verschoor,
Leiden University, Netherlands

Reviewed by:
Daniel Ernst,
Bielefeld University, Germany
Wilfried Kunde,
Julius Maximilian University of Würzburg, Germany

****Correspondence:***
Maurits Adam
maurits.adam@uni-potsdam.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *15 April 2021*
Accepted: *16 July 2021*
Published: *10 August 2021*

Citation:
Adam M, Gumbsch C, Butz MV and Elsner B (2021) The Impact of Action Effects on Infants' Predictive Gaze Shifts for a Non-Human Grasping Action at 7, 11, and 18 Months. Front. Psychol. 12:695550. doi: 10.3389/fpsyg.2021.695550

INTRODUCTION

Humans live in a world that is filled with goal-directed actions: People grasp for objects, use tools for crafting, or extend their hands toward each other to shake them. Per definition, an action is a movement that is performed by an agent in order to obtain a desired goal (Prinz, 1997). Sometimes, it can be crucial to predict the goal of the observed action in order to react accordingly and in a timely manner. In a social environment, this goal prediction is important both in the context of competitive and cooperative situations, which is why its development has been studied extensively over the past decades. Because infants' ability for action prediction is related to their emerging action experience, it is crucial to ask how the developing agentive self supports the processing of non-human action goals.

In developmental psychology, a common measure for infants' ability for action prediction is predictive gaze shifts (e.g., Falck-Ytter et al., 2006; Kanakogi and Itakura, 2011; Ambrosini et al., 2013; Adam et al., 2017). For example, when an infant observes how an agent approaches and grasps a goal object, a predictive gaze shift is coded when the infant's gaze moves from the moving agent to the goal object before the agent arrives there. This predictive gaze behavior has been proposed to reflect attentional mechanisms, where the overt shift of the gaze position from the moving agent to the goal object is preceded by covert shifts of attention to the goal object, given that an agent has been detected (Deubel and Schneider, 1996; Daum and Gredebäck, 2010; Gredebäck and Daum, 2016). Therefore, predictive gaze behavior is a suitable means to investigate infants' ability to identify agents and to process actions as being directed toward a goal.

Starting around 6 to 7 months of age, infants show goal-predictive gaze behavior when observing simple human grasping actions; that is, they shift their gaze to the to-be-reached goal object before the agent arrives at its goal (e.g., Kanakogi and Itakura, 2011; Ambrosini et al., 2013; Adam and Elsner, 2020). A little later, at around 12 months of age, infants even predict the goals of more complex human actions, such as transporting toys into a bucket (Falck-Ytter et al., 2006). It is often suggested that infants' goal-predictive gaze behavior is closely linked to the infants' developing agentive self, acquiring sensorimotor experience with all kinds of actions and their consequences, and infants' ability to perform the observed actions themselves (e.g., Falck-Ytter et al., 2006; Kanakogi and Itakura, 2011; Melzer et al., 2012). This is evidenced by correlations between infants' abilities to perform certain actions and their ability to predict the goal of these actions, as well as by studies showing that infants struggle to predict the goal of actions performed by non-human agents or of actions they are not yet able to perform themselves (e.g., Falck-Ytter et al., 2006; Kanakogi and Itakura, 2011; Cannon and Woodward, 2012; Adam et al., 2017). Additional support comes from looking-time research, in which infants' attribution of goal-directedness to an observed action was measured *post-hoc*, that is, after the action goal had been completed. For example, from 6 months on, infants attribute goals to grasping actions by human hands, but not when a hand touches a goal object with its back, which is an unfamiliar action, or when the grasping action is performed by a mechanical claw, which is an unfamiliar agent (e.g., Woodward, 1998, 1999). Furthermore, at 3 months of age, infants' own production of actions was reported to have a larger impact on infants' goal attribution than have simple observations of the same actions without own production (Gerson and Woodward, 2014a). This suggests that own agentive experience is especially crucial for subsequent action processing, which is also supported by computational models and social developmental data (e.g., Pavlova, 2012; Butz and Kutter, 2017).

The aim of the present study was to investigate 7-, 11-, and 18-month-old infants' predictive gaze behavior during the observation of simple grasping actions performed by an unfamiliar mechanical claw. We expected that predictive gaze behavior will develop across the age groups, presumably due to the infants' increasing prior knowledge about the observed action from own sensorimotor experience and from observing others. Additionally, we studied whether the production of a salient action effect, as a potential agency cue (e.g., Bíró and Leslie, 2007), influences infants' predictive gaze behavior. Recent research suggests that infants are able to predict the goal of actions by non-human agents, as long as these agents exhibit certain behavioral agency cues, such as self-propelled movement, equifinality of goal achievement, or the ability to produce salient action effects (e.g., Bíró, 2013; Adam and Elsner, 2018). For example, Adam and Elsner (2018) presented 11-month-olds with videos of a mechanical claw approaching a toy on a linear path. Remarkably, infants showed goal-predictive gaze shifts not only when the claw showed all three agency cues but also when the claw just grasped the toy and lifted it, therefore displaying only one agency cue, that is, the salient action effect of lifting the toy, which was additionally marked by a sound. This suggests that the action effect was especially important for the infants to predict the observed agent's goal (see Bíró et al., 2014, for similar findings regarding the importance of action effects). In another condition, the claw just grasped the toy and then froze in place. In this case, infants showed tracking gaze behavior; that is, they looked at the claw until it reached the goal. These results are in line with ideomotor accounts proposing that actions are primarily represented by the effects they elicit, which highlights the crucial role action effects play for infants' ability to predict the goal of an observed action event (e.g., Prinz, 1997; Elsner and Hommel, 2001).

Gumbsch et al. (2021) developed a generative, event-predictive computational model that successfully modeled the development of infants' gaze behavior when observing human and non-human agents performing goal-directed actions. The Cognitive Action Prediction Model in Infants (CAPRI) proposes that infants generate internal probabilistic generative models of observed action events and transitions between events. These internal models are suggested to develop through infants' sensorimotor interaction with the environment (e.g., when infants repeatedly grasp for and interact with objects or observe others doing so). Based on the free energy minimization formalism (Friston, 2010; Friston et al., 2015), during action generation and observation, CAPRI actively infers gaze behavior *via* the objective to minimize uncertainty about the probabilistically inferred ongoing and upcoming interactions. Critically, the involved learned, generative, and event-predictive models (Zacks et al., 2007; Butz, 2016; Butz et al., 2021) segment the continuous sensorimotor experiences into event and event-transition encodings, thus enabling deeper considerations about the upcoming events. As a result, predictive gaze behavior developed when CAPRI was trained on object interaction events – in this case not considering differences between observing or executing actions. With hardly any knowledge about grasping actions, the model tracked the moving hand to minimize uncertainty and to gain information about its future position. While learning from the accumulating experience with grasping events, predictive gaze shifts developed, because CAPRI aims at minimizing the uncertainty about whether, when, and how the hand is going to grasp the target object (Gumbsch et al., 2021).

Similarly, Elsner and Adam (2020) argued that actions can be seen as events that are cognitively stored as feature bundles (Hommel et al., 2001; Zacks et al., 2007; Butz, 2016). *Via* an interaction between bottom-up and top-down processes, infants' generation of predictive gaze behavior is suggested to be based on three essential steps. First, bottom-up features of the ongoing, yet incomplete, action have to be perceived and processed. These features include, for example, the agent‘s appearance or the kinematics of the movement toward the goal. Second, this bottom-up information is mapped onto stored action-event schemata, that is, cognitive action representations (Elsner and Hommel, 2001; Zacks et al., 2007; Butz, 2016). Generally, schemata represent organized knowledge that describes different concepts, such as situations or events (Schützwohl, 1998). Therefore, an action-event schema, for example, encodes information acquired through experience as an agentive self, in the form of sensorimotor feature nodes connected by associations of various strength. Schemata are typically used to properly comprehend current input or to predict future input, and therefore, schemata are constantly tested against their compatibility with the observed situation (Schützwohl, 1998). Consequently, the associative network underlying a specific schema is updated frequently, and previously learned associations are adjusted based on new learning experiences. In the case of action-event schemata, the number of feature nodes and the strength of the associations increase upon each performance or observation of an action. For example, action-event schemata encode that when a hand moves toward an object, typically a salient action effect follows once the hand closes-in on that object. Third, when sufficient action experience is available, the perception of bottom-up information about the agent, potential goal object, and start of the movement triggers the inference of a "reaching" action-event schema. This then routes the anticipation of an upcoming salient action effect upon reaching the object, which leads to predictive gaze behavior because the active inference process strives to decrease anticipated effect uncertainty (Elsner and Adam, 2020; Gumbsch et al., 2021).

According to these model considerations, as long as infants have only little to no experience with an observed action or agent, they should not be able to predict the action goal. Instead, tracking the unfolding bottom-up information helps infants to understand the ongoing action, thereby adding feature nodes and strengthening the associations between them in the developing action-event schema. Infants normally gather experience about agents that display various agency cues (e.g., Bíró and Leslie, 2007; Bíró, 2013), and this perceivable bottom-up information appears to be stored in action-event schemata. With accumulating experience, the mere perception of the agent's features or the initial state of the action event becomes sufficient to activate the action-event schema, enabling successful goal predictions (e.g., Elsner and Adam, 2020). Following this idea, when unfamiliar agents, such as mechanical claws, display one or more agency cues, corresponding event schemata (linked to the agency cue) can become activated. As a result, the unfamiliar agent or its observable features, such as its appearance, may become associated with the event schemata. In subsequent trials, these top-down influences then allow for predictive gaze behavior even for the unfamiliar agent.

This is in line with looking-time research showing that infants at 6 months need to see more agency cues than older infants at 9 or 12 months in order to attribute a goal to the action of a mechanical claw (Bíró and Leslie, 2007). Moreover, when 9-month-olds were presented with a situation that suggested that a mechanical claw was about to act goal-directedly, infants' EEG response showed patterns of goal identification (Southgate and Begus, 2013). Finally, in eye-tracking studies, adults showed goal-predictive gaze shifts for unusual hand actions or for grasping by a mechanical claw, even in the absence of any additional agency cues (Kanakogi and Itakura, 2011; Adam et al., 2017). Taken together, these results illustrate how observers with limited knowledge about the observed action rely on the unfolding bottom-up information, whereas observers with more knowledge can rely on their stored top-down information that they have gathered through prior knowledge or experience with an action event or an agent.

A recent study investigated infants' use of bottom-up- versus top-down information across the first year of life, by repeatedly presenting 6-, 7-, and 11-month-olds with a hand that approached and grasped a goal object, followed either by a salient action effect (e.g., lifting up the object, accompanied by a sound) or by just freezing in place (Adam and Elsner, 2020). At 6 months, infants showed tracking gaze behavior regardless of the salient action effect, confirming the assumed behavior of infants who have just recently accomplished the motor development milestone of visually guided grasping. In contrast, at 11 months, when infants are experienced in grasping, predictive gaze behavior occurred in both conditions. Interestingly, at 7 months, infants were predictive in the human-hand condition only with the salient action effect and did not show predictive gaze behavior when a grasping mechanical claw produced the salient action effect. These results might reflect that 7-month-olds' representations for human grasping actions are still weak and were only activated *via* additional agency cues and when infants observed a human hand. For the claw, however, the 7-month-olds did not yet conceive of action representations that could be activated, and therefore, the agency cue did not lead to predictive gaze behavior. These results highlighted how the developing agentive self in infancy might help to shape infants' ability to predict the goals of observed action events.

The current study aimed at taking this idea a step further and at investigating the assumed role of the developing agentive self and the interplay of top-down and bottom-up information in the context of predictive gaze behavior for a non-human grasping action. If observers, depending on their knowledge about an observed action, indeed rely on either prior top-down knowledge about the observed action or on presented bottom-up information when generating predictive eye movements, we should see similar developmental patterns, albeit at different ages, for familiar agents and for unfamiliar agents. For example, compared to human hands, infants have much less conceptual knowledge about how mechanical claws are able to grasp and manipulate objects, or about how claws can be used as tools. Therefore, we predicted that the results by Adam and Elsner (2020) across different age groups could be replicated with an unfamiliar agent such as a mechanical claw, when the infants are older and have more experience with grasping in general, but also with the use of tools (e.g., McCarty et al., 2001).

Additionally, there should be an age at which infants possess sufficient action knowledge and would predict the goal of a mechanical claw even without any additional agency cues.

Therefore, we recorded 7-, 11-, and 18-month-olds' eye movements, while infants repeatedly watched a video in which a mechanical claw approached and grasped a goal object and then either did or did not produce a salient action effect. We investigated first, at which age infants would use the agency cue to predict the goal of a simple grasping action performed by a mechanical claw. Second, we investigated whether there would be a learning process that manifests itself as faster gaze shifts to the goal across trials. The three age groups were chosen based on prior research: We expected the 7-month-olds to not show predictive gaze shifts regardless of the salient action effect based on 7-month-olds' limited knowledge about both grasping actions and tool-use actions, and on research reporting that 7-month-olds do not predict the goal of a mechanical claw even when it produces salient action effects (McCarty et al., 2001; Adam and Elsner, 2020). We expected the 11-month-olds to be predictive in the condition with the action effect, but to show tracking gaze behavior in the condition without the action effect, because they have more knowledge about grasping actions than the 7-month-olds. In previous studies, 11-month-olds, who still have relatively limited experience with tool-use actions (McCarty et al., 2001), showed predictive gaze behavior when a grasping mechanical claw produced a salient action effect, but tracked a mechanical claw in the absence of additional agency cues (Adam and Elsner, 2018). Finally, we expected the 18-month-olds to show predictive gaze behavior regardless of the action effect, because infants at that age should already have sufficient knowledge about grasping actions and tool use. Specifically, between 14 and 19 months, infants start to engage in successful actions with claw-like tools to obtain distant objects (McCarty et al., 2001). Therefore, at 18 months, the advanced agentive self should enable goal-prediction *via* top-down processes upon perceiving the start of the claw's grasping, even without agency cues.

MATERIALS AND METHODS

Participants

The final sample consisted of forty-two 7-month-olds (M = 6.9, SD = 0.3, range = 6.5–7.5 months, 20 girls), forty-one 11-month-olds (M = 10.9, SD = 0.3, range = 10.5–11.5 months, 21 girls), and forty-one 18-month-olds (M = 18.0, SD = 0.3, range = 17.5–18.6 months, 22 girls). An additional 9, 8, and 4 participants, respectively, were tested but had to be excluded because they did not contribute enough valid data (criteria see below). The participants were randomly assigned to either the action-effect condition (7-month-olds: n = 21; 11-month-olds: n = 20; and 18-month-olds: n = 19) or the no-action-effect condition (7-month-olds: n = 21; 11-month-olds: n = 21; and 18-month-olds: n = 22). The parents and their children were recruited from a database where parents can sign up their child to participate in studies in the babylab. Participants mostly came from middle-class families in a small German city. During their stay at the laboratory, parents signed informed consent and received 7.50 € as well as a certificate with a photograph of their child as reimbursement. This study was approved by the Ethics Committee of the University of Potsdam.

Stimuli, Apparatus, and Procedure

Participants were presented with 12 repetitions of a video showing how a claw approached and interacted with a toy. In both experimental conditions, the first part of the video was identical: The videos showed the surface of a gray table filmed from the side in front of a gray background with a toy sitting on the table at screen center (see **Figure 1**). After approximately 1,000 ms, a claw that was painted with a light color entered the scene from the right side of the screen, approached the toy on a linear path, and grasped it (duration approx. 2,140 ms). In the action-effect condition, the claw then lifted the toy up, accompanied by a sound, and put the toy back on the table (duration approx. 2000 ms). Based on prior research, the addition of the sound was not expected to influence infants' predictive gaze behavior (Adam et al., 2017). Then, the screen froze and the scene was presented for another 3,870 ms until the video ended. In the no-action-effect condition, immediately after the claw grasped the toy, the screen froze for approximately 5,870 ms until the video ended. Thus, both videos were identical in length and had a total running time of about 9,010 ms. Attention-getter videos (e.g., a bouncing ball or a waving hand) were presented in between stimulus videos in order to redirect the participants' gaze to the screen.

Gaze behavior was recorded with an SMI RED 250 mobile eye tracker mounted to a 22-inch screen. The sampling rate was 250 Hz, and the screen resolution was 1,680 by 1,050 pixels. During the experiment, participants sat on their caregiver's laps in front of the screen, approximately 60 cm away from the eye tracker. Caregivers had no prior knowledge about the contents of the stimuli or the purpose of the study and were instructed to only interact with their child in case she needed soothing. The experiment started with a 5-point calibration and with manual point acceptance. The calibration stimulus was an animated picture of a pulsating circle in front of a gray background. After successful calibration, the experiment started with a total runtime of about 2.5 min.

Data Handling

In both conditions, we used the same areas of interest (AOIs) to analyze participants' gaze behavior (see Falck-Ytter et al., 2006; Kanakogi and Itakura, 2011; Adam et al., 2016, for similar criteria): a static AOI for the goal object and a moving AOI for the claw. Gaze-arrival times were calculated by subtracting the time when participants first fixated the goal AOI from the time when the claw entered the goal AOI. Gaze-arrival times above the value of 0 ms were considered predictive, gaze-arrival times around 0 ms were considered tracking, and gaze-arrival times below 0 ms were considered reactive. A trial was valid when participants first fixated the claw AOI for at least 200 ms before they fixated the goal AOI. Using this criterion of 200 ms (see Gredebäck and Melinder, 2010; Kanakogi and Itakura, 2011; Henrichs et al., 2012;

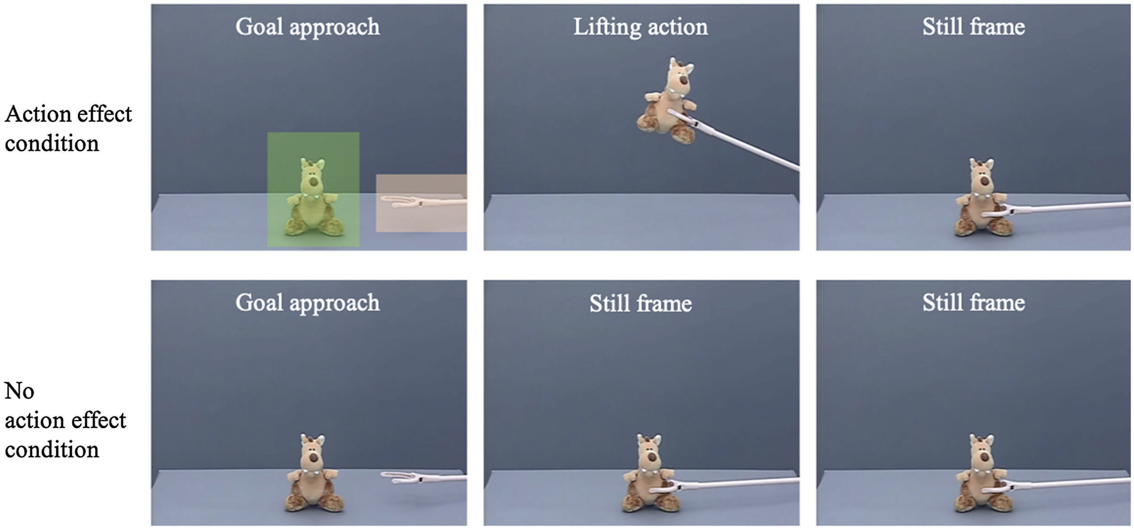


FIGURE 1 | Still frames of the stimulus videos in the action-effect condition (upper row) and the no-action-effect condition (lower row). The two squares in the first picture depict the areas of interest (AOIs) used for data analysis. The squares were not visible during the experiment.

Adam et al., 2016, for a similar use of this criterion) ensured first, that the infants had indeed at least shortly attended to the moving agent and therefore to the movement part of the action. Second, it ensured that infants who just looked at the goal object throughout the action were not included in the analyses, because these sticky fixations would not tell us whether the infant in that particular trial had been predictive. Additionally, values of −1,000 ms or below were classified as invalid. The first trial was excluded from our analyses, because the experimental manipulation of the action effect only occurred at the end of the first video. Participants needed to have at least two valid trials among the analyzed trials 2–12 to be included in our analyses, a criterion that has been used in studies with infants around 6 months of age (Kanakogi and Itakura, 2011; Gredebäck et al., 2018; Adam and Elsner, 2020). The gaze-arrival times in the valid trials were then averaged for every participant to create a mean gaze-arrival time. On average, the 7-month-olds contributed significantly more trials in the action-effect condition (n = 7.8, SD = 3.4) than in the no-action-effect condition (n = 5.1, SD = 2.6), $t(40) = -2.92$, $p < 0.01$, r = 0.4. Both the 11-month-olds (action-effect condition: n = 8.4, SD = 3.1; no-action-effect condition: n = 8.2, SD = 2.6; $t(39) = -0.23$, p = 0.82, r = 0.04) and the 18-month-olds (action-effect condition: n = 9.2, SD = 2.1; no-action-effect condition: n = 9.4, SD = 2.1; $t(39)$ = 0.31, p = 0.76, r = 0.05) contributed a similar number of valid trials in both conditions. In neither age group, there was a significant correlation between the number of valid trials and the mean gaze-arrival time, all $ps > 0.28$.

To test our hypotheses, we conducted ANOVAs and Bonferroni-corrected independent-samples t-tests in order to compare the mean gaze-arrival times as a function of the between-subjects factors age group and condition. We also performed one-sample t-tests against the threshold of 0 ms for every subgroup to classify the gaze behavior as predictive, tracking, or reactive. In case of a null result, we also included BF_{01}, indicating the Bayes factor in favor of the H0 over H1 with values between 0 and 3 representing anecdotal evidence, between 3 and 10 representing moderate evidence, and >10 representing strong evidence. Additionally, we used exploratory regression analyses with linear, logarithmic, and quadratic curve fitting in every subgroup to investigate potential changes of the mean gaze-arrival times across trials 2 to 12. When one of the functions for the gaze-arrival times yielded a significant fit, we also performed exploratory regression analyses with the same functions on infants' fixation times on the claw AOI and on the goal AOI. Linear, logarithmic, and quadratic curve fitting was chosen because significant fits for these types of curves have commonly been reported in prior research (e.g., Henrichs et al., 2012; Adam et al., 2017; Adam and Elsner, 2020).

RESULTS

The ANOVA on mean gaze-arrival time with age group (7 months vs. 11 months vs. 18 months) and condition (action effect vs. no action effect) as between-subjects factors yielded a significant main effect of age group, $F(2,118) = 17.0$, $p < 0.001$, η^2 =0.22, a significant main effect of condition, $F(1,118) = 13.5$, $p < 0.001$, η^2 =0.10, and a significant interaction between age group and condition, $F(2,118) = 3.2$, $p < 0.05$, η^2 =0.05 (see **Figure 2**). Regarding the main effect of the age group, post-hoc independent-samples t-tests (Bonferroni corrected with $\alpha = 0.016$) revealed that mean gaze-arrival times did not differ between the 11- and 18-month-olds, $t(80) = -1.36$, p = 0.18, r = 0.2. However, both the 11-month-olds, $t(81) = -3.83$, $p < 0.001$, r = 0.4, and the 18-month-olds, $t(81) = -5.16$, $p < 0.001$, r = 0.5, had significantly faster mean gaze-arrival times than the 7-month-olds. The main effect of condition resulted from faster mean gaze-arrival times in the action-effect than no-action-effect condition. To explore the significant interaction, we compared the mean gaze-arrival times between conditions for each age group separately by independent-samples t-tests (Bonferroni corrected with $\alpha = 0.016$). Significantly faster mean gaze-arrival times in the action-effect than no-action-effect condition occurred in both the 7-month-olds, $t(40) = -3.65$, $p < 0.001$, $r = 0.5$, and the 11-month-olds, $t(39) = -2.74$, $p < 0.01$, $r = 0.4$, but not in the 18-month-olds, $t(39) = -0.12$, $p = 0.91$, $r = 0.02$.

The one-sample t-tests against the threshold of 0 ms confirmed our expectations: The 7-month-olds' gaze behavior was reactive in the no-action-effect condition, $t(20) = -5.74$, $p < 0.001$, $r = 0.8$, and tracking in the action-effect condition, $t(20) = -0.05$, $p = 0.96$, $r = 0.01$, $BF_{01} = 6$. The 11-month-olds were tracking in the no-action-effect condition, $t(20) = -0.37$, $p = 0.72$, $r = 0.08$, $BF_{01} = 5.6$, and predictive in the action-effect condition $t(19) = 2.96$, $p < 0.01$, $r = 0.6$. Finally, the 18-month-olds were predictive in both the no-action-effect condition, $t(21) = 2.9$, $p < 0.01$, $r = 0.5$, and the action-effect condition, $t(18) = 2.8$, $p < 0.05$, $r = 0.6$.

Regarding potential learning effects, the exploratory regression analyses on mean gaze-arrival times across trials 2–12 in the action-effect condition revealed a significant fit for a logarithmic function for the 7-month-olds ($y = 137.25\ln(x) - 241.55$, $R^2\ adj = 0.44$, $F(1,9) = 8.97$, $p < 0.05$) and a significant fit for a quadratic function for the 11-month-olds ($y = 227.96 + 179.37x - 11.98x^2$, R^2adj $= 0.66$, $F(2,8) = 10.85$, $p < 0.01$). In all other conditions and age groups, the analyses did not yield significant fits, all $ps > 0.06$ (see **Figure 3**). These results indicate that in the action-effect condition, the mean gaze-arrival times of the 7-month-olds got rapidly faster across the first trials, albeit still with mean gaze-arrival times below or at 0 ms, and the mean gaze-arrival times of the 11-month-olds got faster across the first half of the trials, but then slightly decelerated toward the end, always staying above 0 ms. The 18-month-olds' gaze-arrival times generally stayed in the predictive value range above 0 ms across trials in both conditions. Additional exploratory regression analyses on 7- and 11-month-olds' fixation times on the claw AOI and the goal AOI across trials 2–12 in the action-effect condition yielded a significant fit for a quadratic function for the 7-month-olds' fixation times on the claw ($y = 1068.06 - 10x - 2.46x^2$, $R^2adj = 0.86$, $F(2,8) = 32.24$, $p < 0.001$), as well as a significant fit for a logarithmic function for the 11-month-olds' fixation times on the claw ($y = -179.74\ln(x) + 1250.84$, $R^2adj = 0.78$, $F(1,9) = 35.39$, $p < 0.001$), and a significant fit for a linear function for the 11-month-olds' fixation times on the goal ($y = 2530.74 - 44.92x$, $R^2adj = 0.42$, $F(1,9) = 8.34$, $p < 0.05$). These results show that across trials, the 7-month-olds in the action-effect condition looked less at the claw, and the 11-month-olds in the action-effect condition looked less at the claw and at the goal object (see **Figure 3**. For information on fixation times across trials for all six groups, see **Supplementary Figure 1**).

DISCUSSION

The aim of the present study was to investigate the impact of the agency cue of producing a salient action effect in interplay with the developing agentive self on 7-, 11-, and 18-month-olds' goal-predictive gaze shifts during the observation of a non-human grasping action. We investigated at which age and in which conditions infants would be able to produce predictive gaze behavior, and we also looked at potential learning effects across trials. Fitting to our expectations, we found no predictive gaze behavior regardless of the salient action effect in the 7-month-olds, predictive gaze behavior when the salient action effect was presented, but tracking gaze behavior when the salient action effect was not presented in the 11-month-olds, and predictive gaze behavior regardless of the salient action effect in the

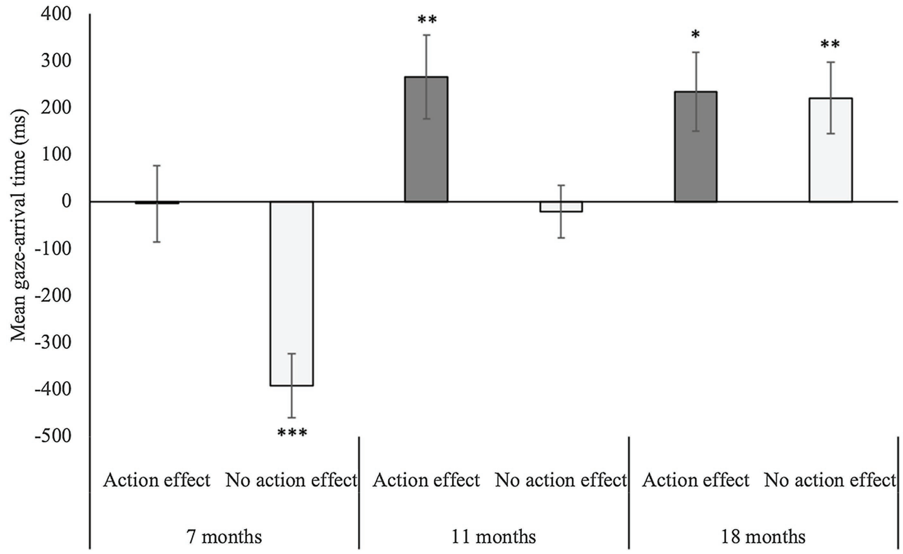


FIGURE 2 | Mean gaze-arrival times for the 7-, 11-, and 18-month-olds in the action-effect and the no-action-effect condition. Positive and negative values represent mean gaze-arrival times before and after the claw arrived at the goal AOI. Error bars represent standard-errors, and the asterisks mark mean gaze-arrival times significantly different from 0 ms. $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

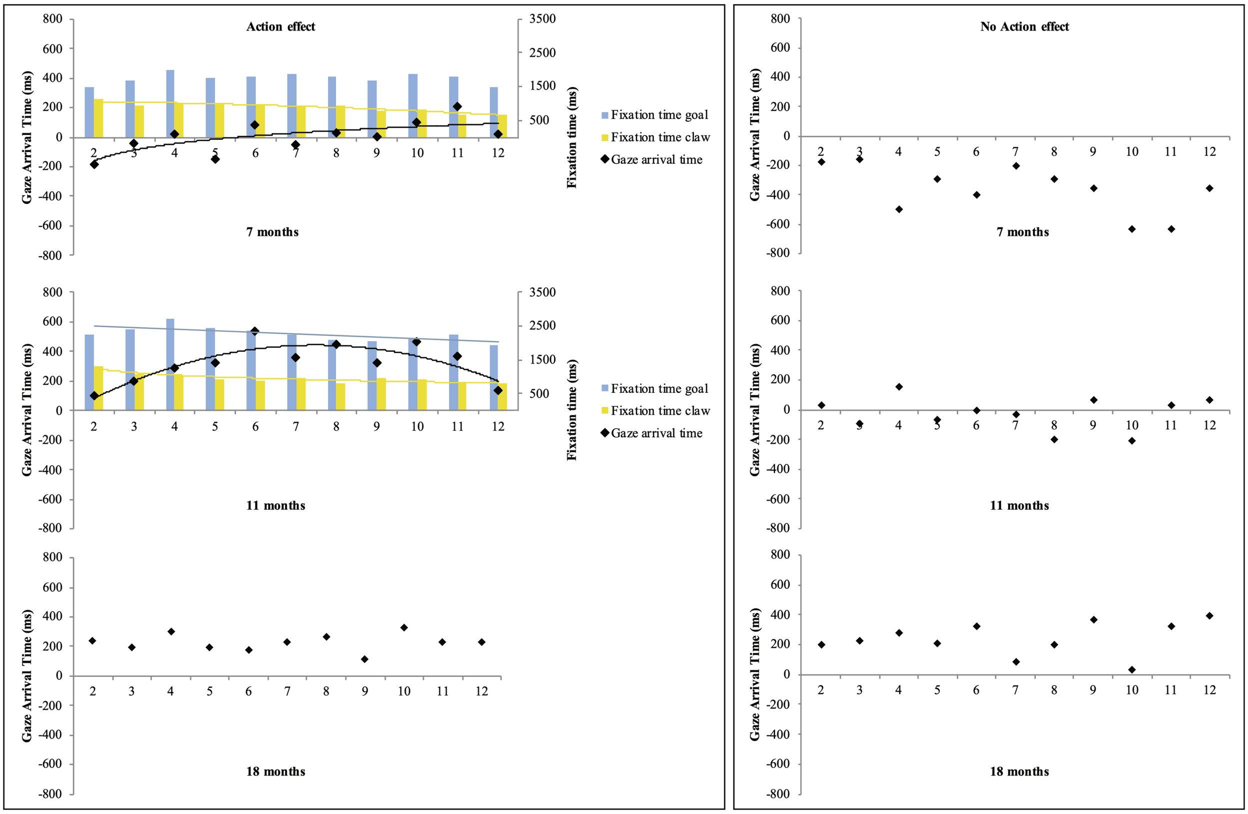


FIGURE 3 | Mean gaze-arrival times (black dots) and fixation times on the goal AOI (blue bars) and the claw AOI (yellow bars) across trials 2–12 for the 7-, 11-, and 18-month-olds in the action-effect and no-action-effect condition. Positive and negative values represent mean gaze-arrival times before and after the claw arrived at the goal AOI. The curves represent the significant fit for the regression functions (linear, logarithmic, or quadratic) with most explained variance.

18-month-olds. This result pattern replicates previous findings in the context of human actions for mechanical actions and shows how similar patterns show up later during development for non-human compared to human actions, which fits the slightly later occurring developmental milestone of tool use compared to grasping (Adam and Elsner, 2020). Additionally, regarding gaze behavior across trials, in the action-effect condition, we found a significant fit for a logarithmic function for the 7-month-olds and a significant for a quadratic function for the 11-month-olds, indicating systematic changes of gaze behavior across trials: The 7-month-olds showed increasing mean gaze-arrival times across the course of the experiment, whereas the 11-month-olds showed increasing mean gaze-arrival times in the first half of the experiment, but decreasing mean gaze-arrival times in the second half (while still staying in the predictive value range).

First, our findings from the 7-month-olds replicate prior research by showing that even in the presence of a salient action effect, infants at this age do not use this information in order to predict the goal of the mechanical claw (Adam and Elsner, 2020). Additionally, the present study expands these findings by showing that although infants' gaze behavior on average was not predictive in the action-effect condition, mean gaze-arrival times in this condition were still significantly faster (i.e., classified as tracking) than in the no-action-effect condition (i.e., classified as reactive). This shows that the agency cue had an impact on the 7-month-olds' gaze behavior, but that ultimately, infants still did not arrive with their gaze at the goal object ahead of time. Further, regression analyses revealed that the 7-month-olds in the action-effect condition showed rapidly increasing mean gaze-arrival times across the first trials. This implies that observing the action effect might have triggered the 7-month-olds' (still weak) action knowledge to a certain degree, and that across trials, the infants indeed used the bottom-up information in the form of the action effect to produce faster gaze shifts toward the end of the experiment. It is interesting to note that the 7-month-olds' gaze behavior for the grasping claw in the action-effect condition was more comparable to the gaze behavior of 6-month-olds, not 7-month-olds, for a grasping hand exhibiting an action effect (Elsner and Adam, 2020), and that in the present no-action-effect condition, the 7-month-olds' mean gaze-arrival times were even lower than that of the 6-month-olds for the grasping hand without an action effect. These results fit the idea that due to the later developing developmental milestone of tool-use compared to grasping (e.g., McCarty et al., 2001), the internal models for mechanical claws also develop later than the internal models for human hands. Therefore, based on the assumptions of the CAPRI model (Gumbsch et al., 2021) and Elsner and Adam (2020), we conclude that the 7-month-olds did not yet have strong action-event schemata for actions performed by mechanical claws, which resulted in tracking gaze behavior or in the case of the no-action-effect condition, even reactive gaze behavior, to maximize information gain and to minimize uncertainty by closely observing the agent and its movement. However, it is possible that the 7-month-olds would be able to predict the claw's action goal when more agency cues were provided, for example, when the claw moved biologically or was self-propelled (e.g., Premack, 1990; Baron-Cohen, 1994; Bíró and Leslie, 2007).

Second, the present findings replicated that 11-month-olds show predictive gaze behavior when a mechanical claw displays agency cues, such as the production of a salient action effect (Adam et al., 2017; Adam and Elsner, 2018). Additionally, regression analyses revealed that for the 11-month-olds, the observation of the action effect resulted in increasing mean gaze-arrival times (in an overall predictive value range) across the first half of the experiment and in a slight decrease (though still in the predictive value range) across the second half. In contrast, in the no-action-effect condition, the 11-month-olds' gaze followed the claw to the goal, with no systematic change across trials. These results are in line with our expectations, confirming that seeing the grasping claw and action effect probably triggered 11-month-olds' grasping experience as well as their emerging action knowledge about tool use (e.g., McCarty et al., 2001). Furthermore, the 11-month-olds' mean gaze-arrival times in both conditions were strikingly similar to the ones found by Adam and Elsner (2020) at 7 months for a grasping human hand. This further indicates that, when stored action-event schemata are still relatively weak, infants can benefit from the display of agency cues, because the cues exert stronger activation of the stored action-event schemata and a stronger agency attribution to the claw (at 11 months). This activation, in turn, may enable goal prediction for subsequent observations of this action event *via* forward modeling and top-down processes (Elsner and Adam, 2020; Gumbsch et al., 2021).

Third, the findings from the 18-month-olds revealed predictive gaze behavior in both conditions. Thus, this study is the first to show that infants at 18 months are able to predict the goal of an ongoing grasping action of a mechanical claw regardless of the salient action effect, that is, in the absence of any additional bottom-up information. Additionally, regression analyses revealed no systematic change of gaze behavior across trials, because the 18-month-olds already started out with predictive mean gaze-arrival times in the first trials. These results indicate that 18-month-olds had already built up strong internal models and strong action-event schemata with regard to grasping and tool-use actions, which they could use as top-down information during the initial observation of the claw, the potential goal object, and the start of the goal approach. At 18 months of age, infants are already quite apt at simple tool-use actions to retrieve a distant object (starting around 14 months of age; McCarty et al., 2001). Based on this, we would possibly find similar results already at an earlier age. However, we chose to study 18-month-olds because infants' ability to produce an action does not instantly guarantee that they would also be able to predict the goal during observation of this action (e.g., Gredebäck et al., 2018; Adam and Elsner, 2020).

The results from the regression analyses for the 7- and 11-month-olds do not match prior findings in which no learning effects were reported for 7-month-olds (Adam and Elsner, 2020) and in which 11-month-olds were reported to show rapidly faster mean gaze-arrival times in the predictive value range across the first trials, but no decreasing mean gaze-arrival times in the second half of the experiment (Adam et al., 2017). Here, it needs to be noted that infants' limited attention span allows for only a very limited number of trials, providing only a weak basis for the analysis of learning effects. For example, the regression analyses on 7- and 11-month-olds' fixation times to the claw AOI and the goal AOI in the action-effect condition revealed that across trials, the 7-month-olds looked less at the claw, and the 11-month-olds looked less at both the claw and the goal object, indicating a decreasing interest in the presented stimuli over the course of the experiment. These results fit to prior research indicating that infants' looking times tend to decrease when a stimulus is repeated multiple times (e.g., Woodward, 1998, 1999). Therefore, it does not come as a surprise that in studies on infants goal-predictive gaze behavior, the results on learning effects across trials are generally unstable and seem to occur unsystematically, even when they are measured with similar stimuli (e.g., Henrichs et al., 2012). Therefore, interpretations of these findings have to be made with caution, and further systematic research on the factors driving learning effects during action observation is needed. For example, it remains unclear whether there is a systematic relation between fixation times on the stimulus display across trials and the corresponding mean gaze-arrival times.

Admittedly, the present findings are ambiguous about whether infants' gaze behavior directly relied on infants' experience with or knowledge about the observed action, or on general cognitive maturation. The role of general maturation processes seems to be supported by the fact that infants' general ability to disengage their gaze from an interesting stimulus improves across the first year of life (Elsabbagh et al., 2013). However, predictive gaze behavior differed as a function of producing a salient action effect in 7-month-olds for a human hand, and in 11-month-olds for a mechanical claw (Adam and Elsner, 2020), which cannot be explained solely by general cognitive maturation processes. Training studies in which infants learn to perform novel actions could be used to disentangle these factors by investigating the impact of systematic manipulation of such learning experience on predictive gaze behavior. For example, a short training session in which infants were encouraged to actively engage in a novel action altered infants' subsequent looking times during observation of that action, indicating changed attribution of goal-directedness (Sommerville et al., 2005; Woodward, 2009; Gerson and Woodward, 2014b). However, the effect of training sessions on infants‘ predictive gaze behavior still needs to be investigated in more detail.

An alternative explanation of our results across the age groups could be that the increasing gaze-arrival times merely reflect the increasing size of infants' functional visual field (e.g., Hullemann and Olivers, 2017). Based on this idea, older (but not younger) infants could have detected the goal object *via* peripheral vision, which in turn triggered an early gaze shift, without any involvement of action processing or the activation of action-event schemata. However, this cannot explain why same-aged infants (i.e., the 7- and 11-month-olds), with functional visual fields matured to a certain size, exhibited significantly higher gaze-arrival times in the action-effect-condition than in the no action-effect-condition. Additionally, Adam and Elsner (2020) found the same pattern of gaze behavior for observations of a grasping hand in younger age groups with a comparably less developed functional visual field. Therefore, although general maturation processes regarding infants' cognition may play a role, action-related cognitive processing has to be in place in order to fully account for our findings.

Another alternative interpretation may be that infants' predictive gaze behavior is not specific to the observation of goal-directed actions performed by agents, but is instead elicited by associative learning of the objects' movements. That is, infants may have shifted their gaze to the goal object because they had learned that "when object A (claw) touches object B (goal object), object B will start moving". For the action-effect condition, we cannot fully exclude impacts of such general learning mechanisms of simple associations between moving objects. However, these mechanisms fail to explain why infants' predictive gaze behavior varies with the familiarity of the observed agent (e.g., Falck-Ytter et al., 2006; Kanakogi and Itakura, 2011; Cannon and Woodward, 2012; Adam et al., 2016). For example, 7-month-olds showed predictive gaze shifts for an effect-producing grasping human hand (Adam and Elsner, 2020), but in the present study did not predict the goal of an almost identical action of a claw. In addition, infants' predictive gaze behavior depends on specific features of the "agent" and of the movement, in particular on cues that signal agency (e.g., Bíró, 2013). Therefore, we take our findings to reflect infants' cognitive processing of observed actions rather than simple associative learning of regularities in the movements of random objects.

Taken together, our results expand the previous work on infants' goal-predictive gaze behavior in the context of human hands to simple actions performed by a non-human agent. Framed according to the theoretical model by Elsner and Adam (2020) and the CAPRI model (Gumbsch et al., 2021), at 7 months, infants' stored action representations are probably still too weak to enable predictive gaze behavior, even in the presence of the agency cue of producing a salient action effect (Bíró and Leslie, 2007). At 11 months of age, infants' stored action representations are still weak, but strong enough to be activated by some observations of the production of a salient action effect during the first trials, which enables goal prediction upon observing the action's start in subsequent trials. Finally, at 18 months, infants' stored action representations are strong enough to be activated already for the first action observations, even in the absence of any additional agency cues. Therefore, these results provide further evidence for the role of the developing agentive self and the interplay between bottom-up and top-down information during the observation of goal-directed actions and illustrate how the shifted developmental courses of this behavior follow infants' motor development and acquired action experience with mechanical agents compared to human agents (Csibra, 2007; Southgate, 2013; Elsner and Adam, 2020; Gumbsch et al., 2021). Future research should further investigate the specific role of infants' agentive experience with the observed action by applying training paradigms, which would shed more light on the interplay of perceptual bottom-up information and experience-based top-down processes underlying the developmental course of infants' goal-predictive gaze behavior.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the University of Potsdam. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MA wrote the manuscript. All authors were involved in the design of the study, data collection, data analyses, and contributed to the manuscript intellectually.

FUNDING

This research was funded by the German Research Foundation (DFG) within Priority-Program SPP 2134 – project "Development of the agentive self" (BU 1335/11–1, EL 253/8–1).

ACKNOWLEDGMENTS

We would like to thank the research assistants who helped to collect the data, and thank all the infants and their parents for taking their time to participate in this study. Martin Butz is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – project number 390727645.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.695550/full#supplementary-material

REFERENCES

Adam, M., and Elsner, B. (2018). Action effects foster 11-month-olds' prediction of action goals for a non-human agent. *Infant Behav. Dev.* 53, 49–55. doi: 10.1016/j.infbeh.2018.09.002

Adam, M., and Elsner, B. (2020). The impact of salient action effects on 6-, 7-, and 11-month-olds' goal-predictive gaze shifts for a human grasping action. *PLoS One* 15:e0240165. doi: 10.1371/journal.pone.0240165

Adam, M., Reitenbach, I., and Elsner, B. (2017). Agency-cues and 11-month-olds' and adults' anticipation of action goals. *Cogn. Dev.* 43, 37–48. doi: 10.1016/j.cogdev.2017.02.008

Adam, M., Reitenbach, I., Papenmeier, F., Gredebäck, G., Elsner, C., and Elsner, B. (2016). Goal saliency boosts infants' action predictions for human manual actions, but not for mechanical claws. *Infant Behav. Dev.* 44, 29–37. doi: 10.1016/j.infbeh.2016.05.001

Ambrosini, E., Reddy, V., de Looper, A., Costantini, M., Lopez, B., and Sinigaglia, C. (2013). Looking ahead: anticipatory gaze and motor ability in infancy. *PLoS One* 8:e67916. doi: 10.1371/journal.pone.0067916

Baron-Cohen, S. (1994). How to build a baby that can read minds: cognitive mechanisms in mind-reading. *Cah. Psychol. Cogn.* 13, 513–552.

Bíró, S. (2013). The role of the efficiency of novel actions in infants' goal anticipation. *J. Exp. Child Psychol.* 116, 415–427. doi: 10.1016/j.jecp.2012.09.011

Bíró, S., and Leslie, A. (2007). Infants' perception of goal-directed actions development through cue-based bootstrapping. *Dev. Sci.* 10, 379–398. doi: 10.1111/j.1467-7687.2006.00544.x

Bíró, S., Verschoor, S., Coalter, E., and Leslie, A. M. (2014). Outcome producing potential influences twelve-month-olds' interpretation of a novel action as goal-directed. *Infant Behav. Dev.* 37, 729–738. doi: 10.1016/j.infbeh.2014.09.004

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

Butz, M. V., Achimova, A., Bilkey, D., and Knott, A. (2021). Event-predictive cognition: a root for conceptual human thought. *Top. Cogn. Sci.* 13, 10–24. doi: 10.1111/tops.12522

Butz, M. V., and Kutter, E. F. (2017). *How the Mind Comes Into Being: Introducing Cognitive Science From a Functional and Computational Perspective*. Oxford, UK: Oxford University Press.

Cannon, E. N., and Woodward, A. L. (2012). Infants generate goal-based action predictions. *Dev. Sci.* 15, 292–298. doi: 10.1111/j.1467-7687.2011.01127.x

Csibra, G. (2007). "Action mirroring and action understanding: an alternative account," in *Sensorimotor Foundations of Higher Cognition: Attention and Performance*. eds. P. Haggard, Y. Rossetto and M. Kawato (Oxford: Oxford University Press), 435–459.

Daum, M. M., and Gredebäck, G. (2010). The development of grasping comprehension in infancy: covert shifts of attention caused by referential actions. *Exp. Brain Res.* 208, 297–307. doi: 10.1007/s00221-010-2479-9

Deubel, H., and Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vis. Res.* 36, 1827–1837. doi: 10.1016/0042-6989(95)00294-4

Elsabbagh, M., Fernandes, J., Webb, S. J., Dawson, G., Charman, T., and Johnson, M. H. (2013). Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood. *Biol. Psychiatry* 74, 189–194. doi: 10.1016/j.biopsych.2012.11.030

Elsner, B., and Adam, M. (2020). Infants' goal prediction for simple action events: The role of experience and agency cues. *Top. Cogn. Sci.* 13, 45–62. doi: 10.1111/tops.12494

Elsner, B., and Hommel, B. (2001). Effect anticipation and action control. *J. Exp. Psychol.* 27, 229–240. doi: 10.1037/0096-1523.27.1.229

Falck-Ytter, T., Gredebäck, G., and von Hofsten, C. (2006). Infants predict other people's action goals. *Nat. Neurosci.* 9, 878–879. doi: 10.1038/nn1729

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

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

Gerson, S. A., and Woodward, A. L. (2014a). Learning from their own actions: the unique effect of producing actions on infants' action understanding. *Child Dev.* 85, 264–277. doi: 10.1111/cdev.12115

Gerson, S. A., and Woodward, A. L. (2014b). The joint role of trained, untrained, and observed actions at the origins of goal recognition. *Infant Behav. Dev.* 37, 94–104. doi: 10.1016/j.infbeh.2013.12.013

Gredebäck, G., and Daum, M. M. (2016). The microstructure of action perception in infancy: decomposing the temporal structure of social information processing. *Child Dev. Perspect.* 9, 79–83. doi: 10.1111/cdep.12109

Gredebäck, G., Lindskog, M., Juvrud, J. C., Green, D., and Marciszko, C. (2018). Action prediction allows hypothesis testing via internal forward models at 6 months of age. *Front. Psychol.* 9:290. doi: 10.3389/fpsyg.2018.00290

Gredebäck, G., and Melinder, A. (2010). Infants' understanding of everyday social interactions: a dual process account. *Cognition* 114, 197–206. doi: 10.1016/j.cognition.2009.09.004

Gumbsch, C., Adam, M., Elsner, B., and Butz, M. V. (2021). Emergent goal-anticipatory gaze in infants via event-predictive learning and inference. PsyArXiv [Preprint]. doi:10.31234/osf.io/9g8uj

Henrichs, I., Elsner, C., Elsner, B., and Gredebäck, G. (2012). Goal salience affects infants' goal-directed gaze shifts. *Front. Psychol.* 3:391. doi: 10.3389/fpsyg.2012.00391

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–937. doi: 10.1017/S0140525X01000103

Hullemann, J., and Olivers, C. N. (2017). The impending demise of the item in visual search. *Behav. Brain Sci.* 40, 1–69. doi: 10.1017/S0140525X15002794

Kanakogi, Y., and Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nat. Commun.* 2:341. doi: 10.1038/ncomms1342

McCarty, M. E., Clifton, R. K., and Collard, R. R. (2001). The beginnings of tool use by infants and toddlers. *Infancy* 2, 233–256. doi: 10.1207/S15327078IN0202_8

Melzer, A., Prinz, W., and Daum, M. M. (2012). Production and perception of contralateral reaching: a close link by 12 months of age. *Infant Behav. Dev.* 35, 570–579. doi: 10.1016/j.infbeh.2012.05.003

Pavlova, M. A. (2012). Biological motion processing as a hallmark of social cognition. *Cereb. Cortex* 22, 981–995. doi: 10.1093/cercor/bhr156

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition* 36, 1–16. doi: 10.1016/0010-0277(90)90051-K

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154. doi: 10.1080/713752551

Schützwohl, A. (1998). Surprise and schema strength. *J. Exp. Psychol. Learn.* 24, 1182–1199. doi: 10.1037/0278-7393.24.5.1182

Sommerville, J. A., Woodward, A. L., and Needham, A. (2005). Action experience alters 3-month-olds infants' perception of others' actions. *Cognition* 96, B1–B11. doi: 10.1016/j.cognition.2004.07.004

Southgate, V. (2013). Do infants provide evidence that the mirror system is involved in action understanding? *Conscious. Cogn.* 22, 1114–1121. doi: 10.1016/j.concog.2013.04.008

Southgate, V., and Begus, K. (2013). Motor activation during the prediction of nonexecutable actions in infants. *Psychol. Sci.* 24, 828–835. doi: 10.1177/0956797612459766

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* 69, 1–34. doi: 10.1016/S0010-0277(98)00058-4

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non- purposeful behaviors. *Infant Behav. Dev.* 22, 145–160. doi: 10.1016/S0163-6383(99)00007-7

Woodward, A. L. (2009). Infants' grasp of others' intentions. *Curr. Dir. Psychol. Sci.* 18, 53–57. doi: 10.1111/j.1467-8721.2009.01605.x

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind/brain perspective. *Psychol. Bull.* 133, 273–293. doi: 10.1037/0033-2909.133.2.273

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

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

frontiers in Psychology

REVIEW
published: 16 August 2021
doi: 10.3389/fpsyg.2021.716671

The Embodied Crossmodal Self Forms Language and Interaction: A Computational Cognitive Review

Frank Röder, Ozan Özdemir, Phuong D. H. Nguyen, Stefan Wermter and Manfred Eppe*

Knowledge Technology, Department of Informatics, University of Hamburg, Hamburg, Germany

Human language is inherently embodied and grounded in sensorimotor representations of the self and the world around it. This suggests that the body schema and ideomotor action-effect associations play an important role in language understanding, language generation, and verbal/physical interaction with others. There are computational models that focus purely on non-verbal interaction between humans and robots, and there are computational models for dialog systems that focus only on verbal interaction. However, there is a lack of research that integrates these approaches. We hypothesize that the development of computational models of the self is very appropriate for considering joint verbal and physical interaction. Therefore, they provide the substantial potential to foster the psychological and cognitive understanding of language grounding, and they have significant potential to improve human-robot interaction methods and applications. This review is a first step toward developing models of the self that integrate verbal and non-verbal communication. To this end, we first analyze the relevant findings and mechanisms for language grounding in the psychological and cognitive literature on ideomotor theory. Second, we identify the existing computational methods that implement physical decision-making and verbal interaction. As a result, we outline how the current computational methods can be used to create advanced computational interaction models that integrate language grounding with body schemas and self-representations.

Keywords: embodiment cognition, grounding language, dialog, minimal self, reinforcement learning, developmental psychology, developmental robotics

OPEN ACCESS

Edited by:
Stephan Alexander Verschoor,
Leiden University, Netherlands

Reviewed by:
Felice Cimatti,
University of Calabria, Italy
Brian MacWhinney,
Carnegie Mellon University,
United States

****Correspondence:***
Frank Röder
frank.roeder@uni-hamburg.de

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *29 May 2021*
Accepted: *16 July 2021*
Published: *16 August 2021*

Citation:
Röder F, Özdemir O, Nguyen PDH, Wermter S and Eppe M (2021) The Embodied Crossmodal Self Forms Language and Interaction: A Computational Cognitive Review. Front. Psychol. 12:716671. doi: 10.3389/fpsyg.2021.716671

1. INTRODUCTION

The human species has a unique communication system that involves verbal (e.g., speech) and non-verbal (e.g., gestures, facial expressions, body language) interaction with others. Despite cultural and social differences, participants in a conversation need to share a common conceptual view of the world and their embodied self. This is essential to have a common understanding, avoid misunderstandings, interpret metaphors (Feldman and Narayanan, 2004) (see **Figure 1A**), and for self-other distinction (Schillaci et al., 2013). A common conceptual view of the world is a consequence of the shared commonalities in how conversation partners ground language in their embodied interaction with the world (Barsalou, 2008; Madden et al., 2010). For example, the common conceptual view implies a self-representation that enables humans to solve tasks involving intrinsic spatial reference frames, like the one in **Figure 1B**. But how can humans learn appropriate

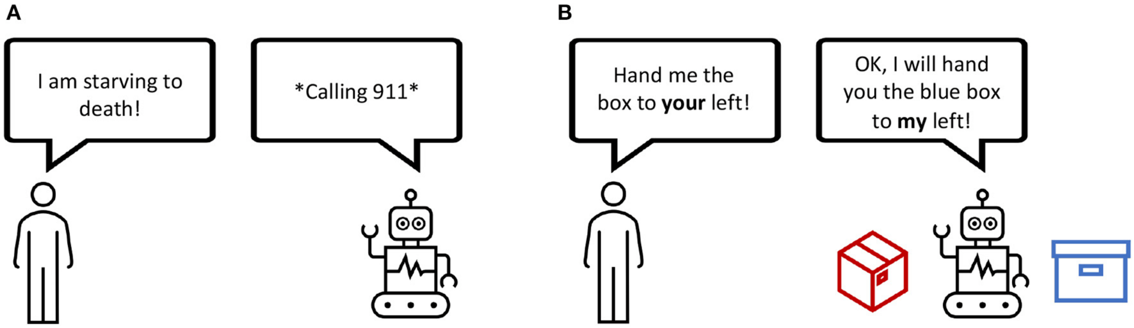

A
I am starving to death!
Calling 911
B
Hand me the box to **your** left!
OK, I will hand you the blue box to **my** left!

FIGURE 1 | (A) Misunderstanding a metaphor, potentially due to the lack of a self-representation. **(B)** A robot using the self as point of reference to understand an instruction.

representations of their body and, consequently, their self? Is the self a unifying principle that combines all the needed ingredients to solve both mentioned examples?

In this review, we will address these questions from an interdisciplinary perspective. Therefore, we will first discuss the cognitive and psychological background for self-representation and embodied language learning. Second, we will align this background with contemporary research in reinforcement learning. Herein, we focus on the cognitive mechanistic aspects of representation learning and behavior. We also appreciate insights from neuroscientific literature (Rizzolatti and Arbib, 1998; Kaplan, 2007; Madden et al., 2010), but we draw only occasional links to maintain a feasible scope for this article, we draw only occasional links to particularly relevant neuroscience background.

1.1. Embodied Language Learning

Human-robot interaction (HRI) is an active field of research where communication via natural language is an essential but also a very challenging component. In the past years, methods utilized machine learning to improve natural language processing (NLP), enabling decent interactions with virtual agents like Siri, Alexa, Cortana, and Google. These improvements are mainly due to utilizing large neural network-based language models (Vaswani et al., 2017; Devlin et al., 2019). However, these systems are limited to disembodied language processing, and therefore, cannot understand how natural language is situated in the physical world. For example, properties such as "heavy" or "hot" cannot be experienced without sensors, and they are important for robots interacting with humans. A robot should understand that hot things can hurt living beings and that not every person can lift heavy objects. There exists research on how robots can technically acquire and understand language through sensorimotor grounding (Steels et al., 2012; Spranger et al., 2014). However, in practice, this is still challenging for current computational models on robots as sensory inputs are imperfect, and natural language is full of ambiguities (see **Figure 1A**). For example, Steels and Loetzsch (2012) present research on how robots can establish new names for objects they see in an environment. They play a grounded naming game with a hardcoded cognitive system and vision, speech recognition, and pointing mechanisms. This is consistent with the concept of decoupling skill learning and language language grounding (Akakzia et al., 2021; Lynch and Sermanet, 2021) that we consider in this article.

To address the problem of imperfect sensors and noisy perception, researchers and engineers often use crossmodal inputs following the notion of the duck test for deductive reasoning: *"If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck."* (Hill et al., 2020; McClelland et al., 2020). Language models, even if showcased as extremely powerful like GPT-3 (Brown et al., 2020), are limited as they cannot make sense of *swimming* or what a *quaking* duck would sound or even look like. To fully understand what swimming and quacking are, an agent requires embodied and situated experiences to ground these concepts. This includes physical interaction with water and, preferably, cross-modal visual and acoustic sensory input to perceive the quacking. In other words, many of the existing language models like GPT-3 perform Natural Language *Processing* (NLP), but they lack the embodied grounding processes required for Natural Language *Understanding* (NLU). As a consequence, to understand language in the context of a dialog and to be able to interact physically with the world via actuators, it is critical to receive embodied multisensory inputs, such as vision, sound, and touch. **Figure 2** illustrates a possible association between the language modality and other modalities (right side) compared to a model that cannot use such grounded connections. Understanding grounded language is critical for acting robots (Tellex et al., 2020) to perform dialog (Bordes et al., 2017) and HRI in general.

Many human skills can be acquired by explanation through language only. However, learning physical skills like a backflip is hard and costly to learn by verbal explanations only because it also benefits from the athletic experience. For example, Christiano et al. (2017) were able to teach an agent to do a backflip via simple feedback akin to basic language only, describing how good the agent is currently performing or what to improve.

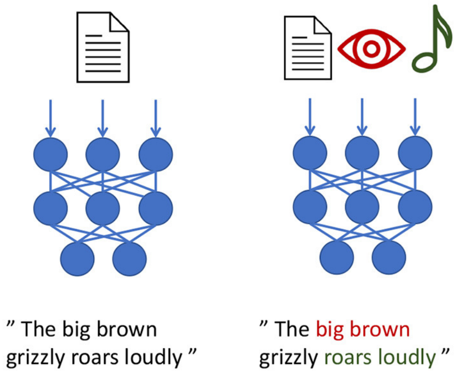

" The big brown grizzly roars loudly "
" The big brown grizzly roars loudly "

FIGURE 2 | We illustrate an example using a neural text-processing model that integrates text only **(left)** and text in combination with vision and sound **(right)**. Possible associations or groundings are highlighted.

The key point is that learning skills through language require embodied concepts that recall motions and postures in context. For example, *"While jumping as high as you can, pull your legs towardz your body and throw yourself to the back; after a full rotation, land on your feet"* presupposes that the skill *"jumping"* is already known. Without such concepts, explaining the execution of a backflip, similarly to the example of Christiano et al. (2017), requires a vast amount of feedback or very detailed guidance to compensate for the lack of knowledge.

In summary, humans leverage embodied concepts built up during their lifetime, with language understanding always tightly connected to knowledge and experiences of the motor system (Fischer and Zwaan, 2008). Specifically, verbal descriptions like *"throwing a ball"* or *"jumping in the air"* excite the relevant parts of the motor cortex that are active for both hearing and executing. Therefore, language acquisition is strongly influenced by embodied experiences and the current context (McClelland et al., 2020).

1.2. Reinforcement Learning and Computational Language Understanding Methods

Reinforcement learning (RL) (Sutton and Barto, 2018) is a cognitively plausible and valuable framework to emulate infant-like learning, exploring the world with a trial-and-error approach based on rewards. RL-based agents are sometimes intrinsically motivated (Forestier et al., 2017; Colas et al., 2020; Akakzia et al., 2021; Hill et al., 2021). They imitate behaviors (Chevalier-Boisvert et al., 2019; Lynch and Sermanet, 2021), use hierarchical abstractions to decompose a complex task into simpler tasks (Oh et al., 2017; Eppe et al., 2019), and some of them can be trained with language to follow instructions (Hermann et al., 2017; Oh et al., 2017; Chaplot et al., 2018; Narasimhan et al., 2018; Chevalier-Boisvert et al., 2019; Hill et al., 2019, 2020, 2021; Jiang et al., 2019; Colas et al., 2020).

Reinforcement learning is also a promising method to implement dialog systems (Shi and Yu, 2018; Saleh et al., 2020) and language-driven interactive RL (Cruz et al., 2015; Chevalier-Boisvert et al., 2019). Commonly, language in RL (Luketina et al., 2019) is either used to provide an instruction (what to do) or to assist the learning of the agent with hints and descriptions (Narasimhan et al., 2018). Other methods describe the agent's environment purely in textual form, e.g., the agent's state in a dialog or text-based game (Côté et al., 2019; Madureira and Schlangen, 2020), which is a common setup for most conversational settings. For example, the simulator ALFWorld (Shridhar et al., 2021) was published with the goal to provide a learning environment where they combine the text-based knowledge obtained in TextWorld (Côté et al., 2019) is combined with visual inputs from ALFRED (Shridhar et al., 2020). Saleh et al. (2020) use hierarchical reinforcement learning (HRL) (Barto and Mahadevan, 2003) in an open-domain dialog, providing results that are comparable with the current state-of-the-art language models (Vaswani et al., 2017). As another example for language-driven RL, consider the research by Jiang et al. (2019), who use simplified language to communicate between a lower and higher layer of a hierarchical RL agent following language instructions.

The recent review by Uc-Cetina et al. (2021) illustrates the applicability of RL in NLP to some extent, such as machine translation, language understanding, and text generation. The authors also suggest considering embodiment (Heinrich et al., 2020), textual domain knowledge, and conversational settings. Bisk et al. (2020) focus further on embodiment and highlight the importance of physical and social context, more precisely, multimodal sensory experiences, to apprehend the coherency of words and actions. In an embodied dialog, the notion of technically combining the world state, i.e., the sensory inputs, with a linguistic state of a dialog, e.g., the context of the last *n* utterances, is crucial. We also see advances in multimodal reinforcement learning (Schillaci et al., 2013; Chaplot et al., 2018; Hill et al., 2019, 2020, 2021), integrating multisensory experience for explainability and improved training performance.

1.3. Scientific Rationale and Contribution of This Review

The work of Eppe et al. (2020) provides a thorough review of the hierarchical concepts for embodied problem-solving, but the authors do not consider language. Another related review about computational models of the self and body schemas has recently been presented by Nguyen et al. (2021). However, the authors do not consider language either. We address this gap by examining the challenges of embodied dialogs (Hahn et al., 2020) in the context of the self, combining the presence of language with other input modalities to learn appropriate hierarchical representations.

For our review, we hypothesize that a disembodied combination of the latest insights in multimodal data processing and language processing is not sufficient to enable full language understanding in dialogs between humans and embodied computational agents like robots. Instead, we hypothesize that an increased focus on the embodied self is important to enable computational agents with true language *understanding* capabilities beyond the mere computational *processing* of language. We investigate this hypothesis by addressing the following research questions:

What are the cognitive components of the self, and why are they important for communication and dialog? Which components have been realized computationally, and how? Which are still missing?

To address these questions, and as our main contribution, we look into recent articles defining the prerequisites of an artificial self (Schillaci et al., 2016; Georgie et al., 2019; Hafner et al., 2020; Nguyen et al., 2021) and relate these prerequisites with verbal and non-verbal dialog methods for computational agents and reinforcement learning. In section 2, we survey the developmental processes of humans to ground language in embodied sensorimotor representations of the self and its surrounding world. In section 3, we summarize existing computational methods that use grounded language to train an agent. In section 4, we address our main hypothesis by summarizing and detailing why the self contains all the components that make robots better language learners and dialog partners. In addition, we provide a blueprint for combining the different existing computational techniques. These results are followed by a brief conclusion in section 5.

2. COGNITIVE AND PSYCHOLOGICAL PERSPECTIVES OF THE COMMUNICATING SELF

The development of the human ability to perform bi-directional language-based dialog is a process over three interleaved stages. The first stage is sensorimotor development, where infants learn to align their perception with their motor skills (Paul et al., 2018) to acquire an understanding of the physical dynamics of their environment. Based on such low-level sensorimotor knowledge acquisition, humans develop embodied mental concepts in a second developmental stage to model their environment in higher-level preverbal conceptual representations (Feldman, 2006; Barsalou, 2008; Frankland and Greene, 2020). Such higher-level concepts are the foundation of language, which emerges with social interaction and communication during the third stage of development (Feldman, 2006; Kiefer and Pulvermüller, 2012). These three stages are not temporally distinct, but they co-develop. For example, verbal interaction demands additional low-level motor skills to produce phonemes using tongue, lips, and diaphragm. And social interaction leads to learning new conceptual representations that describe social interaction, e.g., in meta-communication. In the following, we will summarize the psychological and cognitive foundations of each of these stages.

2.1. Learning Sensorimotor Representations

From the very first month of birth, infants start developing a sense of their own body and its relation to other physical entities, such as objects and other living beings (Nguyen et al., 2021). The representation of their body in space that encodes positional and relational information is called the body schema (Holmes and Spence, 2004; Hoffmann et al., 2010). The body schema, or sense of body, is mainly shaped by proprioception, but visual information and other modalities (Wermter et al., 2009), including sound, vision, pain, and smell, also play a role (Anderson, 1972). The multimodality of the formation of low-level sensorimotor representations is very efficient for humans suffering from a lack of one or more senses. For example, visually impaired humans can build a rich conceptual understanding of words, objects, and the world, even without the visual sense (Nguyen et al., 2021). Generally, the absence of one or more modalities can be compensated by the other modalities, such as touch and sound. Therefore, multisensory integration is crucial for embodied cognition and learning concepts to represent the world.

Ideomotor theory postulates that the physical knowledge about multimodal sensorimotor contingencies is encoded as bi-directional action-effect associations (Shin et al., 2010). This implies that neural structures learn a mapping between actions and effects that enable humans to predict the outcome of actions and external events. The same structures enable humans to select an action based on a *desired* effect, i.e., a goal. The acquisition of ideomotor associations is enabled by observing and interacting with the world, learning principles such as occlusion, solidness, collision, gravity, and other physical events (Baillargeon, 2001).

Developmental psychology suggests that the acquisition of sensorimotor knowledge is guided by several forms of intrinsic motivation, including self-guided play (Sutton-Smith, 2001), curiosity (Oudeyer et al., 2007), repetition, and imitation (Wood et al., 1976; Paulus, 2014). Self-guided play implies that infants conduct their own experiments, e.g., dropping toys to discover forces like gravity, to extend their knowledge about the world and their own capabilities (Sutton-Smith, 2001). This behavior is closely tied to curiosity and active learning: infants often strive to encounter surprising and unpredictable situations to maximize their knowledge about the world (Schwartenbeck et al., 2019). More specifically, Schwartenbeck et al. (2019) state that active learning builds on minimizing the *unexpected uncertainty*, which can be described as the uncertainty about uncertainty. The authors exemplify active learning with a two-armed bandit problem where the reward of using one arm is low, but the agent knows that the probability for the low reward is high. The other arm has a low but unknown probability for a high reward. In this case, an agent will first try to resolve the unexpected uncertainty about the unknown probability for a high reward of the second arm by trying it. In general, it will collect samples of state transitions with a high unexpected uncertainty until it has a good estimate of the uncertainty.

This explorative behavior, however, must be balanced with striving for predictable action-state transitions, as described by the *free energy principle* (Friston, 2009). This principle implies that humans and other acting systems perform an *active inference* behavior and seek to encounter predictable situations. It describes long-term surprise as an upper limit for free energy and states that biological agents strive to minimize the free energy. At first glance, active inference seems to contradict the active learning behavior where agents strive to encounter uncertain and unpredictable situations to maximize their knowledge gain.

However, since active learning seeks to encounter situations with a high *unexpected uncertainty*, i.e., uncertainty about uncertainty, this is in fact very compatible with active inference, which seeks to avoid situations with a high *expected uncertainty*. In other words, active learning is preliminary to active inference because it is required to learn a model about expected uncertainty.

Another form of intrinsic motivation is repetition: Biological agents exhibit behaviors that are not only goal-driven but exclusively conducted for the purpose of repetition to discover multiple possible ways of achieving a goal (Burghardt, 2006). For example, one can think about a child stacking blocks just for the sake of stacking rather than the goal of building a big tower. In the goal-driven case, repetition allows experiencing many ways of achieving the same desired outcome.[1] Acevedo-Valle et al. (2020) point out that intrinsically motivated sensorimotor exploration is also related to imitation. The authors' proposed architecture highlights imitation-based learning of an infant in the pre-linguistic phase, being supervised by an instructor. They consider the simulation of a vocal tract as a comparison to what young infants do to produce vocal sounds when acquiring speech. Most robots do not have a vocal tract, but there exists research on modeling goal-directed behavior where the goal is to produce a certain vowel or syllable (Philippsen, 2021). Here, the authors consider the case of speech acquisition, where goal-directed explorative behavior uses sounds to learn vowels and syllables via *goal babbling* (Philippsen, 2021).

In summary, explorative play and active learning are the main drivers for learning to "know the unknown" (Vygotsky, 1967; Belsky and Most, 1981) and, more specifically, about the effects and uncertainties of actions (Nguyen et al., 2021). However, explorative behavior is balanced with the free energy principle, causing agents to strive for predictable situations. Other drivers of sensorimotor learning are imitation and repetition. Once enough knowledge is acquired, humans and other animals can use their rich conceptual knowledge for one-shot problem-solving (Eppe et al., 2020).

2.2. Formation and Grounding of Preverbal and Abstract Conceptual Representations

Language allows humans to express thought. However, explicit verbal language is not a prerequisite for thought—there exists a preverbal hierarchical system of abstract mental concepts to enable thought (Frankland and Greene, 2020).

2.2.1. Representational Abstraction

The human mind constantly performs inference on multiple layers of representational abstraction (Clark, 2016). The theory of embodied cognition suggests that the higher levels of abstraction emerge from the sensorimotor interaction of the lower levels (Barsalou, 2008; Lakoff and Johnson, 2010; Tani, 2016). Already during the first year of a human's life, sensorimotor abstraction leads to higher-level preverbal concepts that enable problem-solving and the understanding of simple language (Mandler, 2004). These concepts are grounded in sensorimotor experiences and perception, being later on shaped by our acquired language. Cognitive sciences often refer to such preverbal general concepts as *image schemas* (Lakoff and Johnson, 2010; Turner, 2015) or, in a more linguistic context, *semantic frames* (Barsalou, 2008; Gamerschlag et al., 2014).

How exactly such concepts are represented in biological neural structures remains largely unknown. In particular, there is a lack of research concerned with the semantic compositionality of mental concepts. There exists phenomenological research from the cognitive sciences community to model compositional high-level concept formation (Lakoff and Johnson, 2010; Turner, 2015; Eppe et al., 2018). On the other end of the spectrum, there also exists very low-level neuroscientific research showing the compositionality of distributed neural activation patterns via neuroimaging (Haynes et al., 2015). Between these extremes, there is some very interesting work related to binding neurons (Shastri, 1999) that can potentially model semantic role-filler bindings known from cognitive linguistics. The event segmentation theory (EST) is a biologically plausible model to explain action abstraction based on prediction errors (Zacks et al., 2007). However, to the best of our knowledge, no computationally verified and functional unifying theory integrates the cognitive sciences and linguistics perspective on symbolic compositional mental representations with the neuroscientific perspective of representing mental concepts as distributed neural activation patterns.

2.2.2. Abstract Mental Concepts for Language and Creative Thought

Abstract preverbal concepts are not only critical for language acquisition, but they are also very important for creativity (Turner, 2015). For example, consider the metaphorical concepts of files and folders of a computer's operating system: the terminology for these concepts comes from the pre-digital age, originally from non-electronic paper-based files and folders. Blending this terminology with the tree-based algorithmic pointer concepts behind a computer's file system was a creative act that made it possible to align a human's pre-existing conceptual system with new technology and helped to improve the usability of early operating systems like Windows 95. Confalonieri et al. (2015, 2016, 2018) and Eppe et al. (2018) demonstrate the importance of such concept blending with a functional computational model that allows an artificial agent to combine two known concepts to new concepts with emergent useful and aesthetic properties. The authors show how the new blended concepts lead to the creative and serendipitous discovery of lemmas required for mathematical proofs and the automated (re-)discovery of famous chord progressions in jazz music.

2.3. Embodied Language Acquisition

Preverbal and abstract semantic concepts are the basis for language. Since abstract concepts emerge from low-level sensorimotor interaction, the body and environment have a great impact on our thinking and language acquisition (Feldman and Narayanan, 2004). Several studies highlight that hearing or reading language about action and perception activates related areas of the brain, showing that there are neural representations

[1] This idea was recently used to learn robust and diverse behaviors in goal-directed RL (Akakzia et al., 2021; Lynch and Sermanet, 2021).

reflecting an individual's way of performing actions when heard (see the overview by Willems et al., 2010 or the work about the mirror system by Rizzolatti and Arbib, 1998). This is compatible with ideomotor theory (Shin et al., 2010) and mental simulation theory, which claims that humans simulate actions unconsciously within those areas of the brain responsible for motor planning. As a result, there exists an embodied mental semantics (Feldman and Narayanan, 2004; Steels, 2007; Willems et al., 2010), implying that living entities with different kinds of bodies simulate in different ways. For example, consider the difference between right- and left-handed people, using the contrary sides of the premotor cortex.

2.3.1. Language Acquisition as Resolution of Mismatches

Mandler (2004) describes the preverbal phase in infants as dominated by general conceptual knowledge that is in a mismatch with the language we understand and start to use at the age of 9 months. General conceptual knowledge is required to execute goal-directed actions, understand spatial relationships and the difference between objects and animals. The conceptual knowledge is also important to derive non-trivial intentions of conversation partners (Trott et al., 2016). Consequently, when language becomes more important during a toddler's early life, there is a need to compensate for the mismatch between the rich self-acquired conceptual knowledge and the words used to describe the world. For example, toddlers would assign the word dog to a fox since they do not yet have the language to differentiate them more precisely (Mandler, 2004). Similar to machine learning models with the objective of classifying foxes, wolves, and specific breeds of dogs distinctively, a child would pay at some point closer attention to the details if the appearance is different, but the describing word stays the same (Mandler, 2004). One can also think about the attributes mentioned, like *black cat*, *red car*, or *big dog*, to accentuate a specific property, helping with the mapping of words to organize categories (Waxman and Markow, 1995). Mainly using a mixture of receptive language and producing words and simple sentences allows them to learn about things being said to and about them. Especially parents often explain to their children what they are doing, allowing them to learn word mappings to actions and objects nearly automatically, known as perceptual learning (Mandler, 2004). There is also a lot of imitation involved, e.g., replicating actions of social partners, repeating perceived utterances, or recalling sentences in a specific context.

There are still open questions at which point in time infants are capable of learning specific differences, especially those that are hard to grasp, like varieties between similar-looking plants that are not that frequently experienced in their daily life (Mandler, 2004).

2.3.2. Toward Narrative, Egocentric, and Goal-Directed Language

When the first form of language is learned, infants tend to use egocentric speech, where they narrate their own activities (Piaget, 1926). Even though they do not have fully learned fluent language like adults, they use their present concepts and actively reinforce their speech in their own doing. This is different from babbling from an earlier stage, where the overall learning goal is to explore and correct their internal motor model of speech production with respect to adult language heard (see section 2.1). Furthermore, after infants learn a first basic corpus of language, they start using it to describe their intrinsically motivated goals. This can happen by just saying the word "arm" to tell their caregiver that they want to be picked up or by issuing more complex multi-word sentences of the form "I want X," where the "I" reflects an emerging concept of the self (Georgie et al., 2019). Such goal-directed utterances to caregivers are among the first language-based communication situations.

2.3.3. The Self and Communication

Language is very effective when it comes to communicating with other humans. The efficiency stems from the compositional structure of natural language. Most natural languages build on a finite vocabulary in the order of magnitude of 100,000 to 200,000 actively used words that can be composed to express an intractable number of different sentences and meanings. Our acquired knowledge about grammar, syntax, and semantics enables us to understand most of these compositions, even if we have never heard them before. For example, you may never have heard the sentence "She sneezed the napkin off the table.", but your knowledge about English grammar enables you to correctly understand it. This demonstrates that language is an important cognitive tool to convey meaning (Mirolli and Parisi, 2011; Colas et al., 2020; Eppe and Oudeyer, 2021). However, the self described in recent literature (Schillaci et al., 2016; Hafner et al., 2020; Nguyen et al., 2021) is also important for embodied dialog. The self builds upon the actor's capabilities to sense its own body and the environment. It is, therefore, characterized by the response to actions and predictions of the internal model (Schillaci et al., 2016; Hafner et al., 2020). Grounded language in the context of the self refers to the context of these senses. For example, the phrase *"Hand me the box to your left."* (see **Figure 1B**) requires the robot to classify and detect the desired object (Matuszek et al., 2012) that is next to itself. Once the sentence is understood, a sequence of motor controls needs to be executed to fulfill the instruction. While the language already contains important contextual information, such that it is a box and not another object, which requires different balancing and grasping, the clue *"next to you"* suggests the object be in reachable distance, also described as peripersonal space (Nguyen et al., 2021) with respect to the self. The executed actions are conditioned on the initial instruction of handing over the bottle. The theory about the mirror system by Rizzolatti and Arbib (1998) hightlights the linkage between language and action representations (Wermter et al., 2009): Humans can merely recognize the intent of others by observing their behavior, e.g., if someone is approaching another person offensively. Intention recognition, however, plays a core role in communication and dialogs. We build on this neuroscientific perspective to underpin our claim that a self- and other-manifold is essential for embodied dialogs.

Current computational methods cannot effectively learn a theory of mind with the concepts of *you* and *me*. Therefore, they fail to learn robust and general behaviors. We suppose that this gap is due to a lack of understanding of "the self" (Hafner et al., 2020), and how it is defined in the context of "the other." Specifically, we suggest that a self-other projection model is critical for empathy and a theory of mind to map an observed other agent, along with its semantic properties and relations, to the self and its semantic properties and relations.

In the following section, we will address this gap by investigating the computational language acquisition models that exist and summarize how they relate to the cognitive, psychological, and neurological perspectives on the communicative self.

3. COMPUTATIONAL METHODS

Current advances in neural language modeling accelerated the research progress in many NLP tasks (Vaswani et al., 2017; Devlin et al., 2019). Successful pre-trained one-shot models like GPT-3 (Brown et al., 2020) have many useful applications. Remarkable results were presented with the recently introduced successor version of GPT-3, named DALL-E (Ramesh et al., 2021), which learns visual-linguistic representations that align textual with image inputs to generate, based on text descriptions, samples of new pictures, showing up compositional conceptualization. For example, the sentence *"a red table in shape of a pentagon"* lets the model generate samples of red pentagon-shaped tables based on its learned multimodal representations. However, models like GPT-3 and DALL-E consider only disembodied language learning without any sensorimotor grounding because, unlike robots, they cannot physically interact with the world. Insights for grounded language learning in robotics (Heinrich et al., 2020) with sequential decision-making settings (Akakzia et al., 2021; Lynch and Sermanet, 2021) and embodied cognition (Feldman and Narayanan, 2004; Fischer and Zwaan, 2008) accentuate the need for embodied grounding. This includes physical interaction and multiple sensory modalities to develop systems that understand language more like humans (Anderson, 1972; Wermter et al., 2009; McClelland et al., 2020). Additional prerequisites for modeling a communicative self requires curiosity, body representations, and predictive processes (Hafner et al., 2020; Eppe and Oudeyer, 2021). In reinforcement learning, there is a body of research (Pathak et al., 2017; Dean et al., 2020; Nguyen et al., 2020; Röder et al., 2020), containing these components. However, to the best of our knowledge, these prerequisites have not yet been combined with language and the self in mind. Overall, there is a lack of research methods that regard the self in the area of RL, explicitly making use of language in embodied dialogs (Hahn et al., 2020). This section reviews methods that partly satisfy the requirements but still miss at least one of the desired components. Furthermore, we provide an outlook on what needs to be recombined or is missing to learn self-other representations in embodied dialogs.

3.1. Formal Background

Reinforcement learning (Sutton and Barto, 2018) is based on a Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$, where $\mathcal{S}$ is the space of all possible states, $\mathcal{A}$ the space of all possible actions, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, \infty)$ the transition probability function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, and $\gamma \in [0, 1)$ is the discount factor. The transition function represents a probability density of transitioning to a following state $s' \in \mathcal{S}$, when executing action $a \in \mathcal{A}$, being in state $s \in \mathcal{S}$. The reward function describes the immediate real-valued reward obtained when transitioning to the next state. The overall objective is to find a policy π that selects actions, $\pi(a_t|s_t)$, to maximize the expected discounted reward $\sum_{t=1}^{T} \mathbb{E}_\pi \left[\gamma^t R(s_t, a_t)\right]$ for every time step t.

3.1.1. RL and Imitation Learning

The definition of the MDP, as mentioned earlier, also applies to the framework of imitation learning (IL) (Atkeson and Schaal, 1997; Lynch and Sermanet, 2021), where the learner only has access to a sequence of state-action pairs $(s_{1:T}, a_{1:T})$ of an expert—hence the optimal or suboptimal policy—without knowing the reward function R.

3.1.2. Language as Goal

In this review, we consider papers that also augment this setup with a set of goals $\mathcal{G}$ and condition the action-selection of the policy based on the present state and goal, $\pi(a_t|s_t, g_t)$, also named as goal-conditioned RL (Oh et al., 2017; Chaplot et al., 2018; Chevalier-Boisvert et al., 2019; Jiang et al., 2019; Colas et al., 2020; Röder et al., 2020; Akakzia et al., 2021; Lynch and Sermanet, 2021). One way of integrating language into the augmented MDP, is to learn a mapping from language to goal, $m(l_t) \to g_t$. Another approach is to provide extra input to the policy or concatenate and extend the dialog state as a combination of language and world state, $s_t = \left[s_t^{world}, s_t^{dialog}\right]$. However, these are technical questions that we do not further consider within this article.

3.2. Recent Advances in Reinforcement Learning With Language

Modeling language occurrences in a simulated environment is not obvious to implement, and using human-annotated linguistic training data is usually inefficient and costly. It is also a very specific design decision, how complex the sentences and how limited the vocabulary of words used to train the agent are (see section 3.3).

The review of Luketina et al. (2019) provides an overview of the recent progress of language-processing RL agents where researchers explore possibilities of integrating neuro-plausible principles, such as intrinsic motivation (Forestier et al., 2017; Colas et al., 2020), to foster language learning. Many approaches benefit from mapping instructions to action sequences (Branavan et al., 2010; Misra et al., 2017), latent plans (Lynch and Sermanet, 2021), semantic goals (Akakzia et al., 2021), and internal abstractions (Jiang et al., 2019). In section 3.3, we further examine the possibilities of providing language data to artificial agents that learn from sparse rewards as successfully presented by recent approaches (Luketina et al., 2019; Dean et al., 2020; Akakzia et al., 2021; Lynch and Sermanet, 2021). We see a trend of detaching from the traditional MDP formulation and integration imitation-based (Lynch and Sermanet, 2021) and self-supervised methods

(Akakzia et al., 2021) into a learning framework to autonomously acquire motor skills and language understanding with minimal human intervention. We draw inspiration from the intrinsically motivated learning of infants, like mentioned in section 2, based on a cognitive and developmental perspective.

3.2.1. Dataset-Driven RL Methods

Generally, methods make use of sparse goal annotations (Akakzia et al., 2021; Lynch and Sermanet, 2021) or generate scene-dependent descriptions (Narasimhan et al., 2018; Hill et al., 2021) and instructions (Hermann et al., 2017; Oh et al., 2017; Chaplot et al., 2018; Chevalier-Boisvert et al., 2019). Such methods often build on a previously collected fixed dataset. Therefore, most language-conditioned and language-assisted agents are limited in these settings as they do not reveal behavioral diversity, sticking to a poor set of discovered solutions. This is a problem for embodied agents in dialogs and HRI, with potential uncertainties and inaccuracies coming with dynamics of the physical world. Furthermore, many do not consider all the available modalities to build rich and robust representations, including self-representation (Nguyen et al., 2020). Recent work shows that RL with language needs another type of benchmarking, similar to supervised learning, evaluating the agent on unseen tasks, objects, and instructions (Hill et al., 2020). Otherwise, one could not prove the generalizability of learned feature representations that encode concepts and meanings that are relevant. Especially for our case, we consider an embodied conversational setup with an agent and a human communicating, where having a self-other representation is beneficial if not crucial (see **Figure 1B**).

3.2.2. Adding Dynamic Data and Language Grounding

Using datasets only to train RL-based dialog agents creates limitations. However, datasets can be used for pre-training when a basic understanding of language is necessary to solve a certain task. They can also be augmented with other data, such as demonstrations and pre-trained word embeddings. This can also be combined with other learning methods, such as inverse RL.

Interesting perspectives in this direction are covered in the work of Luketina et al. (2019): The authors consider language-conditioned RL, where language processing is inevitable to fulfill a task because either the state space or action space contains language. A sequence of instructions needs to be followed, telling the agent what to do or which goal to accomplish. The authors argue that following high-level instructions has a strong connection to hierarchical RL (HRL) (Oh et al., 2017; Jiang et al., 2019), decomposing the overall dialog into a sequence of subtasks (Röder et al., 2020).

Another approach presented in the same study (Luketina et al., 2019) is to infer the reward function from the present instructions, especially where no external reward is available, but a set of demonstrations is present. A suitable strategy in such a case is inverse RL (Ng and Russell, 2000). An optimal or suboptimal policy trajectory is used to reconstruct the underlying reward function R as the origin of the demonstration policy's behavior. Unlike behavior cloning, as the simplest form of imitation learning, a goal-achievement reward function could be learned (Colas et al., 2020), which could also be helpful for intrinsically motivated- and transfer learning.

Next, Luketina et al. (2019) consider language-assisted RL, which is also partly related to language-conditioned RL, where language eases the learning and is not required to solve a task. Here, language is descriptive and contains assisting clues for the agent, e.g., "be careful with the delicate plates" (as additional hint before the agent tries to pick them up) or "to open a door, it needs to be unlocked with a key" (the agent is facing a door and is stuck or randomly tries to find a solution). This setting requires the agent to retrieve the relevant information for a given context, where a grounded language understanding is inevitable.

Lynch and Sermanet (2021) show that combining imitation learning with pre-trained word embeddings enables zero-shot learning. Approaching problems with pre-trained models like BERT from Devlin et al. (2019) can circumvent the effort to train so-called *"tabula rasa"* RL agents (Luketina et al., 2019), that is, agents that need to learn language and sensorimotor control simultaneously from scratch. Conclusively, language is a vehicle for transfer learning, as it encodes world knowledge distilled from large text corpora (Devlin et al., 2019; Brown et al., 2020). We believe that language in RL (Luketina et al., 2019) should focus on aligning its sensorimotor representations, learning from multisensory inputs (Hill et al., 2021; Ramesh et al., 2021) that exploit and ground the present compositional and hierarchical linguistic concepts.

3.3. Language Data for RL Agents

When infants interact with their caretakers and the world, they receive visual, auditory, and haptic feedback. In addition, they are also exposed to linguistic utterances and speech in the context of this interaction. In machine learning, this corresponds to interactive RL (Cruz et al., 2015). However, as opposed to human infants that can learn from a few examples very efficiently, RL agents require large amounts of interaction data to learn a reasonable behavior. Furthermore, the required presence of a human partner in the training process is still costly and time-consuming. For this review, we consider approaches (1) that can efficiently collect language before training (Chaplot et al., 2018; Narasimhan et al., 2018), (2) that can automatically generate linguistic instructions at training and testing time (Hermann et al., 2017; Chevalier-Boisvert et al., 2019; Jiang et al., 2019; Hill et al., 2020, 2021), and (3) that require only minimal linguistic input for an agent in the learning process (Colas et al., 2020; Akakzia et al., 2021; Lynch and Sermanet, 2021).

3.3.1. Gathering Data in Advance

Approaches that fall into the first category, such as Narasimhan et al. (2018) and Chaplot et al. (2018), gather language data in advance. Narasimhan et al. (2018) utilize Amazon Mechanical Turk (Buhrmester et al., 2011) to collect descriptions of entities (their roles or behaviors) in different game environments—Amazon Mechanical Turk offers a crowdsourcing website where researchers can hire so-called crowd workers to collect large amounts of data easily and rapidly for a particular task. For each game environment, annotators are shown videos of gameplay and asked to describe entities in terms of their role or behavior, whereby a set of descriptions are collected. It is important to note that the annotators are prompted to give descriptive information about the entities rather than instructive information, which may help the agent complete the given task. The agent, in turn, exploits the appropriate set of descriptions in an end-to-end learning process to reach its goal for a given environment. Chaplot et al. (2018), on the other hand, manually create 70 instructions that prompt the agent to navigate in a 3D game environment and find the target object. Each instruction follows the template "Go to the X" where X is an object with its properties such as "green torch," "tall blue object" etc.

3.3.2. Automated Generation of Verbal Instructions

The second category approaches, such as Chevalier-Boisvert et al. (2019) and Jiang et al. (2019), can automatically generate language input during training and testing. Jiang et al. (2019) use the *CLEVR* language engine (Johnson et al., 2017), which programmatically generates scenes of objects and language descriptions/instructions. This also requires the agent to learn a language-conditioned policy in an end-to-end fashion (see section 3.2). In this sparse-reward setting, the authors use *hindsight instruction relabeling* (Jiang et al., 2019) to improve sample efficiency. Chevalier-Boisvert et al. (2019) introduce a synthetic language, the Baby Language, which has a systematic definition with combinatorial properties. Albeit a proper subset of English, the Baby Language has 2.48×10^{19} possible instructions. It has a special grammar based on which synthetic instructions with different actions (pick up, drop, move), colors, objects, and locations (e.g., "move the green ball next to the blue box") can be generated.

3.3.3. Training With Sparse Data

Lynch and Sermanet (2021) and Akakzia et al. (2021) are considered in the third category because they require only very little language data for the agent during the learning process. Lynch and Sermanet (2021) introduce multicontext imitation, which allows flexibility to use paired state-action language data for less than 1% of the examples to train the agent. They pair play data with human language, which they call *hindsight instruction pairing*. They randomly select a robot behavior from play and ask human annotators to describe it with the most suitable instruction, with the question "Which language instruction makes the trajectory optimal?" in their mind. From goal image examples, a paired goal image and language dataset is created that consists of short trajectories paired with unrestricted instructions collected from human annotators. Akakzia et al. (2021) utilize a synthetic social partner that describes the actions of the robotic arm manipulating objects in a simulator.

The first two category methods that we review in this paper do not strictly follow the approach we propose in this work. Many of them integrate the language data directly into the simulation. For our approach, we consider two phases (see **Figure 4**) where data collection is important: *skill learning* and *language grounding*. As a first phase in the *skill learning* (Akakzia et al., 2021), the agent curiously collects data to learn goal-directed behaviors, similar to infants in their preverbal phase (see section 2), shaping their body schema (Nguyen et al., 2020). Subsequently, a social partner or caregiver provides the language to be grounded in the present goal-directed motor skills. Like infants, the agent should align and learn word meanings with the corresponding action effects. We consider a sparse annotation like applied in Lynch and Sermanet (2021) with *hindsight instructions* of $< 1\%$ of demonstrations—proposing the optimal instruction after the fact—or behavior annotations like (Akakzia et al., 2021) with only 10% of episodes as plausible approaches in line with the sparse utterances an infant experiences.

3.4. Decoupling Language Grounding From Skill Learning

We visually summarize our review of research with respect to different approaches used in language-driven RL in **Figure 3**. The figure illustrates the underlying techniques, showing the most overlaps with respect to the categories *multitask*, *hierarchy*, *curiosity*, and *hindsight* in RL. Based on this categorization, we identify two methods that we consider most appropriate to address the research question of this article, namely Lynch and Sermanet (2021) and Akakzia et al. (2021). Among the approaches we discuss here, only these two consider the decoupling of learning skills and grounding language for an embodied robot in a 3D environment. This is important because in order to benefit from insights of preverbal goal-conditioned behavior in human infants (Wood et al., 1976; Mandler, 2004), artificial agents should be able to learn sensorimotor skills without the presence of language right at the beginning of the learning process. For our following discussion, we perform an in-depth analysis of these two methods. Based on the insights from section 2, we split the overall learning into two phases, as shown in **Figure 4**: *skill learning* and *language grounding*.

3.4.1. Skill Learning

The skill learning phase (**Figures 4A,B**) treats the sensorimotor skill learning as (a) learning those skills independently via imagined goals or concepts like self-play and intrinsic motivation or (b) emulating the behaviors of a caregiver via imitation or supervised learning. In the first case (**Figure 4A**), the agent could learn via intrinsically motivated play or mental problem-solving (imagination) to explore possible block configurations (Akakzia et al., 2021). This is similar to how an infant learns by exploring the environment while interacting with the objects around.

In the second case (**Figure 4B**), the agent could learn by imitating the caregiver (Lynch and Sermanet, 2021). Lynch and Sermanet (2021) conducted imitation learning on a dataset of play data. One benefit of play data is the unrestricted setup without solving any particular tasks. In their setup (Lynch and Sermanet, 2021) have a fixed robot arm in front of a desk with buttons, a cupboard, and other objects. The dataset is collected by recording the proprioceptive inputs, images from the camera, and executed motor control. Herein, the agent benefits from a knowledgeable human collecting the data. This yields a dataset of diverse and curious behaviors, including knowledge about object affordances.

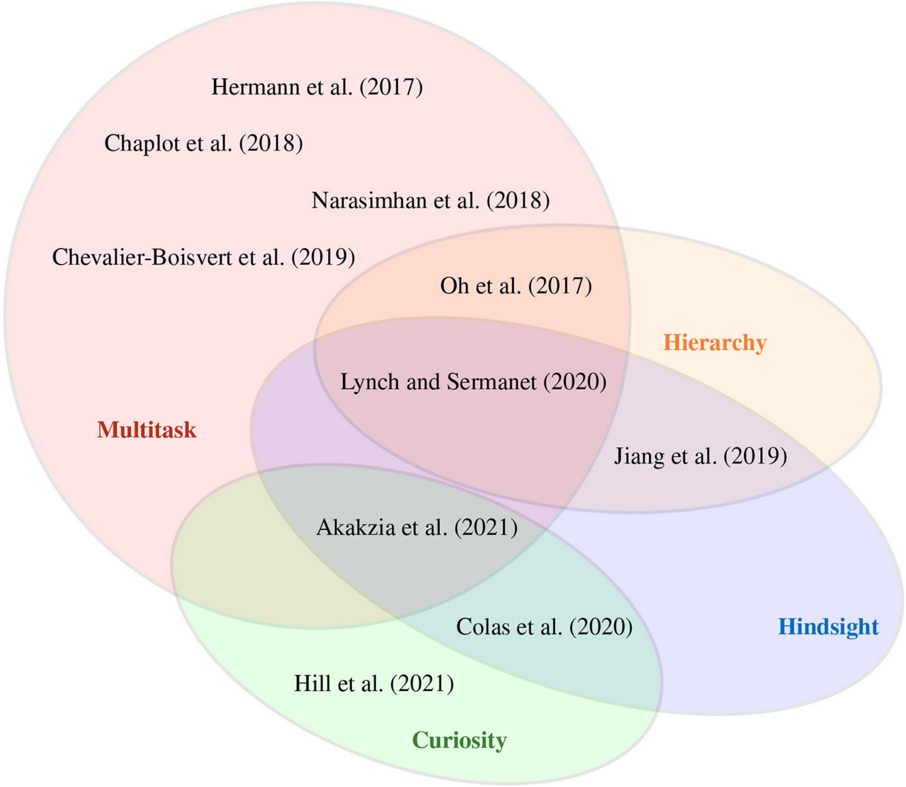


FIGURE 3 | A selection of reinforcement learning methods which we categorize according to their properties. *Multitask* RL involves methods that learn a policy to solve and transfer knowledge between different tasks. *Hindsight* learning allows to create and learn from imagined—(Colas et al., 2020) and relabeled goals (Akakzia et al., 2021). Methods using a *hierarchy* of policies/models are employed for temporal abstractions (Jiang et al., 2019; Lynch and Sermanet, 2021). *Curiosity* serves as an intrinsic signal to utilize self-supervision and overcome sparse extrinsic feedback (Colas et al., 2020; Akakzia et al., 2021; Hill et al., 2021). The methods with the largest overlaps, namely Lynch and Sermanet (2021) and Akakzia et al. (2021), integrate both essential and cognitive plausible mechanism.

3.4.2. Language Grounding

In the second phase (**Figures 4C,D**), learning a grounded language is achieved by providing feedback or instructions. In Akakzia et al. (2021), a social partner—in our case, a caregiver (**Figure 4C**)—provides linguistic feedback, describing the behavior of the agent in hindsight. The social partner provides a description that considers a change in spatial relations between any two objects from the starting configuration to the final in the scene. Language grounding is achieved via a language-conditioned goal generator (LGG) which is implemented as a conditional variational autoencoder (Sohn et al., 2015): given an initial configuration and a description, LGG generates a corresponding final configuration, the goal for the agent to achieve. Resampling from the LGG allows the agent to solve the instruction in different ways, resulting in a diverse behavior (see section 2.1). Similar to Lynch and Sermanet (2021), only a small fraction of the author's dataset is annotated with instructions. These are provided in hindsight: after observing a particular behavior of the agent, the human provides the optimal *"hindsight instruction"* that would evoke this behavior.

Lynch and Sermanet (2021) extend the learning from play (LfP) approach (Lynch et al., 2020) by pairing experienced trajectories with natural language instructions, which they coin as LangLfP. They introduce *multicontext imitation* to train a single policy on both image and language goals. Multicontext imitation refers to training a single policy on shared latent representations of goal image and natural language datasets using image and language encoders. Multicontext imitation endows the approach with the flexibility to use paired state-action language data for less than 1% of the examples to train an agent. Having the ability to learn from sparsely annotated data corresponds with how infants learn in the real world with very little feedback from their caregivers. The trained agent can relate language to low-level perception, perform visual reasoning and solve a complex sequential decision problem. As a result, it can follow non-expert human instructions to perform object manipulation tasks in a row.

Lynch and Sermanet (2021) also exploit a large-scale pre-trained language model (Vaswani et al., 2017; Yang et al., 2020) to encode linguistic input; before feeding the language input to the network, they transfer it to a semantic vector space by using the pre-trained language model as an encoder. In this manner, the approach can handle unseen linguistic inputs such as synonyms, as well as instructions in 16 different languages. We suppose that

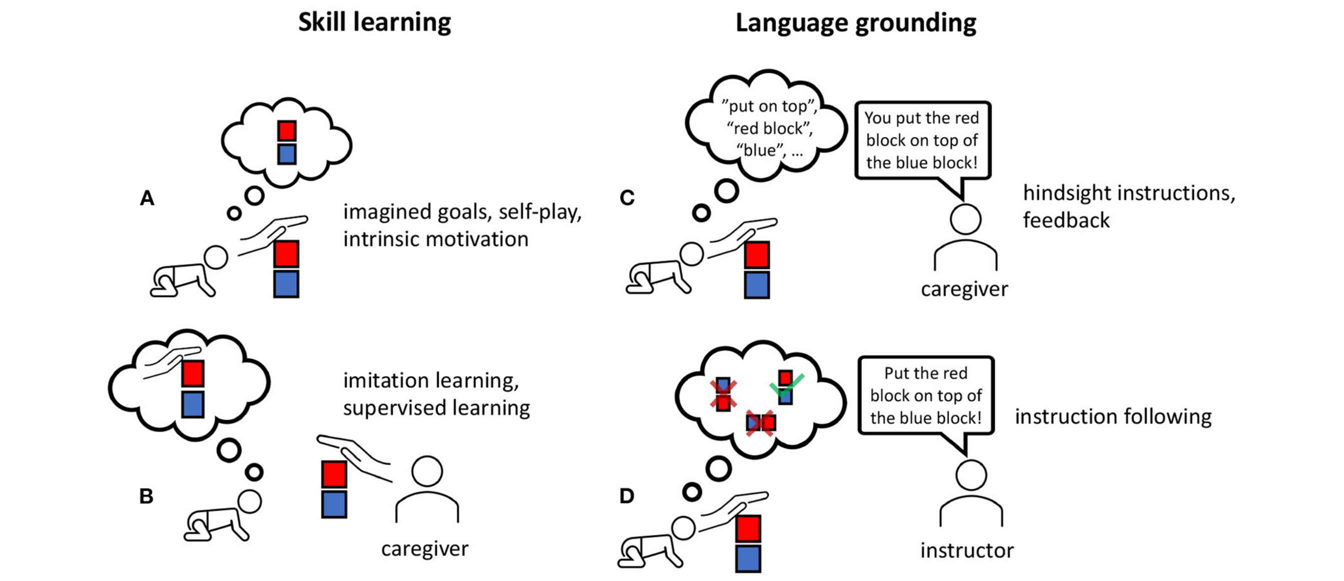

Skill learning
Language grounding
A
imagined goals, self-play, intrinsic motivation
B
imitation learning, supervised learning
caregiver
C
"put on top", "red block", "blue", ...
You put the red block on top of the blue block!
hindsight instructions, feedback
caregiver
D
Put the red block on top of the blue block!
instruction following
instructor

FIGURE 4 | We accentuate the learning phases of current methods (Akakzia et al., 2021; Lynch and Sermanet, 2021) that have a grounded language acquisition by first learning behaviors/skills **(A,B)**—be it via imitation learning or intrinsic motivation—and following this, ground language in actions by receiving instructions or feedback from a caregiver **(C,D)**.

training instruction-following and training dialog are suitable tasks for fine-tuning a pre-trained agent (**Figure 4D**). Moreover, continuing to learn a pre-trained mapping of new objects to concepts appears to be a promising future approach to consider (Hill et al., 2021).

4. THE SELF IN AN EMBODIED DIALOG

In this section, we propose the computational components of an embodied dialog agent, informed by the above analysis of skill learning and language grounding and inspired by the recent work about self-representations of Hafner et al. (2020) and Nguyen et al. (2021).

Naively, testing the capabilities of a language-aware agent could already involve tasks and instructions that specifically strain grounded language knowledge and self-other distinction (see **Figure 1B**). However, we assume that research progress can be accelerated by observing the problem from a perspective of the artificial self (Hafner et al., 2020; Nguyen et al., 2021) rather than disregarding the emerging properties as a side effect. The recent methods introduced in section 3 provide important techniques that implement the required ingredients and are helpful in improving embodied dialogs and HRI applications. Still, we see a lack of methods that combine all of them jointly into one learning architecture.

Current RL methods without language representations can be extended with it (section 3.4.2), as they already include the skill learning phase (section 3.4.1). This is an important feature of RL because skill learning is a necessary prerequisite for language grounding. However, since language grounding is not a necessary prerequisite for skill learning, we conclude that RL-driven physical skill learning is more foundational for embodied dialog agents than disembodied language processing models like GPT-3 (Brown et al., 2020).

In the remainder of this section, we summarize the computational components that are important to develop embodied dialog agents based on self-representations. In addition, we provide references to successful implementations of these components. We subdivide these components into those that are related to predictive processes and those that are related to self-other distinction.

4.1. Predictive Processes and Crossmodal Self-Representations

Many methods compute prediction errors with inverse- and forward models that implement action-effect associations [e.g., Schillaci et al., 2016; Röder et al., 2020 and also neuroscience-related work like (Kaplan, 2007; Kidd and Hayden, 2015)]. At training time, these errors yield a signal for intrinsic motivation, helping to shape and update the body schema and sense of agency (see section 2.1). We see plenty of methods that implement this as curiosity-driven learning (Pathak et al., 2017; Nguyen et al., 2020; Akakzia et al., 2021; Hill et al., 2021). Other researchers model the prediction error not only with the sensory state but based on language. For example, Hermann et al. (2017) and Hill et al. (2021) consider word predictions given the egocentric view of an agent in a 3D environment. Hermann et al. (2017) predict a word at each time step, while a meaningful word of the current instruction serves as a target, e.g., the object *"apple"* given the instruction *"Pick up the red apple."* This auxiliary task helps to shape the agent's representation in learning instruction

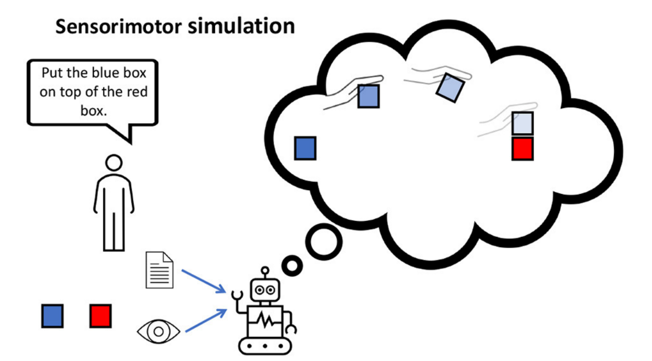


FIGURE 5 | Internal models are capable of mentally simulating possible action trajectories given the visual observation and instruction of stacking the blocks. The longer the simulation horizon, the more uncertain the agent is about its predicted action-effects (illustrated with increasing color transparency).

to word mappings. Hill et al. (2021) compute a surprise score for both vision and language. An episodic memory with a specific language to vision key-mapping, inspired by dual-coding theory (Paivio, 1969), is queried to calculate a language- and vision-based distance as an intrinsic reward. Although this seems to be a promising approach, it is essential to consider some sort of weighting (Hill et al., 2021).

The authors empirically show that the less frequently encountered language is more important than the more frequently changing visual information. However, they are not using an appropriate body representation (Pathak et al., 2017; Nguyen et al., 2020) for the vision encoding to omit the *Noisy-TV Problem* (Burda et al., 2019), which might be the reason for the superior performance when using intrinsic rewards based on language only. Dean et al. (2020) implement an audio-visual association model to employ curiosity-driven exploration by exploiting the associations of two modalities, namely audio and vision.

The approaches above combine crossmodal integration in curiosity-driven and goal-directed learning procedures crucial for intelligent explorative behaviors (Georgie et al., 2019). When evaluating a trained agent, the internal models disclose metrics of surprise where the agent encounters dynamics that are novel or uncertainties with understanding instructions.

Other important computational components for embodied dialog agents include hierarchical abstraction (Eppe et al., 2020) and automatically generated subtasks (Jiang et al., 2019) or latent plans (Lynch and Sermanet, 2021) to abstract away from low-level motor execution, toward higher-level conceptual representations. Abstractions are important because they limit the horizon of predictive processes. For example, in **Figure 5**, we illustrate sensorimotor simulation, using the internal model to unroll a latent (abstract) plan consisting of four steps only. If the same plan was represented in more fine-grained lower-level motor actions, this would lead to many more consecutive simulation steps, resulting in a higher cumulative prediction errors. Also, since predictions become less accurate the farther they are in the future, regenerating plans and subtasks happen more frequently. For example, Lynch and Sermanet (2021) use a hierarchy with a high-level module (plan encoder) to generate a latent plan at the frequency of 1 Hz, while a low-level action module (plan decoder) is executing motor controls at a frequency of 30 Hz. Similarly, the implementation of (Jiang et al., 2019) employs a 2-layer hierarchy that effectively leverages the compositionality of language to solve a task by solving subtasks.

Finally, having access to the agents internal hierarchical predictive state also allows observing metrics such as surprise and uncertainty (e.g., by measuring the prediction error) that expose how strong the sense of body ownership and agency is (Georgie et al., 2019; Hafner et al., 2020).

4.2. Self-Other Distinction

The scenario of **Figure 1B** requires the agent to understand the meaning of self-related words like *you* and other related words like *me*. Georgie et al. (2019) propose that distinguishing self-generated from externally produced sensational actions-effects are inevitable for an artificial self. By dividing the training procedure into two phases (section 3.4), agents learn the required body representations as describe by Georgie et al. (2019), Nguyen et al. (2021), and Hafner et al. (2020). The authors consider motor babbling as an active self-exploration process, starting with self-touch in prenatal development up to toddlerhood. Considering the progression from this early stage, the evolved body ownership and sense of agency define the minimal self (Georgie et al., 2019). We suppose that this stage is covered by our first phase (**Figures 4A,B**), employing motor babbling to train the internal models and motor skills from scratch.

The language-grounding phase (section 3.4.2) exploits the learned behaviors and body representations. This can be performed with a social partner or hindsight instructions to annotate behaviors. With the sense of body ownership developed during the skill learning phase, through minimal prediction error or free energy of inverse- and forward models, the agent can align its motor skills with grounded language. Social-psychological scientists like Mead et al. (2000) postulate the emergence of a self requires a social process based on the social theory of *symbolic interactionism*. However, there are limitations and different perspectives (Aksan et al., 2009) toward social RL (Jaques et al., 2019) and grounded language in a social context (Bisk et al., 2020). We consider these as future work and out of the scope of this article. Nevertheless, according to symbolic interactionism, self-awareness is a kind of reflection and inference of the behavioral observation of others. In other words, the self develops as a generalization of others, putting perception and expectations into the perspective of the social partners or group (Mead et al., 2000). This process allows sharing the same common understanding and thus the same language.

Despite the potential importance of social interaction, our review in section 3 reveals that only Chevalier-Boisvert et al. (2019) contain some sort of interactive partner or teacher that provides linguistic and demonstrative feedback. The authors use a 2D environment and employ a synthetic simplified language (section 3.3). We suggest two possibilities to enhance the integration of a social partner to train a self-aware agent for communication.

The first possibility follows the approach of Chevalier-Boisvert et al. (2019), where the language grounding phase integrates a social partner, caretaker, or teacher. This agent supplies language annotations in hindsight (Akakzia et al., 2021) and, in addition, serves as an embodied entity that provides perceptible demonstrations in combination with language. The second possibility to develop a self for embodied dialog agents is to introduce a third alignment phase (see section 3.4), similarly to the developmental process of section 2.3.3, that involves external crossmodal sensory inputs of a social partner and considers fine-tuning the present motor-linguistic skills of the previous phases (sections 3.4.1 and 3.4.2).

In both cases, the language must explicitly refer to the individuals. Sentences like "You put red on top of the blue" or "I put red on top of blue" are possible examples that allow observing self- and externally generated stimuli in the context of language (McClelland et al., 2020).

5. CONCLUSION

This review contributes to the development of artificial agents for embodied crossmodal dialog. Our main hypothesis is that an explicit self representation is a critical component to enable embodied language understanding, going beyond disembodied language processing as proposed in recent machine learning articles. Reinforcement learning seems particularly suitable, as it allows by definition to discover the environment in a self-explorative manner, similar to an infant shaping its body schema within a self-conducted reinforcement process. Like Lynch and Sermanet (2021) and Akakzia et al. (2021), we suggest splitting the training of an agent into two phases, namely skill learning and language grounding (section 3.4). These two methods are the only ones regarding an embodied robot in a 3D environment and integrate most of the plausible concepts (see section 2 and **Figure 3**) with state-of-the-art performance for complex instruction following. After the skill learning phase, language is grounded in sensorimotor- and body representations, hence in essential parts of the artificial self (Hafner et al., 2020).

As our main result and contribution, we propose and summarize computational components to implement and model an artificial embodied dialog agent in section 4. Here, we highlight self-related components and expand the decoupled two-phased learning to a setting with an embodied social partner.

This approach is underpinned in social-psychological science (Mead et al., 2000) and by recent findings in neurorobotics (Hafner et al., 2020; Nguyen et al., 2021) which emphasize the significance of learning socially with other agents. These benefits arise because self-awareness and natural communication are learned by distinguishing self-generated from external stimuli and being part of social interaction. We believe that explicit self-representations in artificial agents improve robustness, performance, and trust for conversational settings because the emergence of a self is a consequence of low-level interaction with its body and environment (Schillaci et al., 2016; Hafner et al., 2020) and high-level verbal/non-verbal social interactions (Mead et al., 2000).

In this article, we focus primarily on mechanistic cognitive models, but we are also aware of the valuable neuroscientific research that examines the use of the RL framework (Botvinick and Weinstein, 2014), grounded language (Friederici and Singer, 2015; Garagnani and Pulvermüller, 2016), and curiosity (Kaplan, 2007; Kidd and Hayden, 2015). Considering the integration these neuroscientific theories would add a valuable additional dimension to our future research.

A simulation of the self with artificial agents is another beneficial future research direction. For example, we can potentially gain more insights from attention-based mechanisms (Chaplot et al., 2018; Hill et al., 2019), enabling us to visualize the agent's internal state as a kind of gaze following and eye tracking [see Hill et al. (2019), how they visualize the attention weights of different neural network layers when processing language and vision]. Such research paves the ground for measuring and defining neurologically inspired low-level metrics of an artificial agent's self in the future.

AUTHOR CONTRIBUTIONS

FR and ME authored and conceptualized the major parts of this article. OÖ mainly authored and contributed to section 3, revised the manuscript, and was involved in discussions with FR and ME. PN provided feedback for FR to conceptualize the initial outline. SW contributed through active feedback and revisions. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

FR, PN, SW, and ME acknowledge funding by the DFG through the IDEAS (402776968) and LeCAREbot (433323019) projects. OÖ and SW acknowledge support from the German Research Foundation DFG, project CML (TRR 169).

REFERENCES

Acevedo-Valle, J. M., Hafner, V. V., and Angulo, C. (2020). Social reinforcement in artificial prelinguistic development: a study using intrinsically motivated exploration architectures. *IEEE Trans. Cogn. Dev. Syst*. 12, 198–208. doi: 10.1109/TCDS.2018.2883249

Akakzia, A., Colas, C., Oudeyer, P.-Y., Chetouani, M., and Sigaud, O. (2021). "Grounding language to autonomously-acquired skills via goal generation," in *International Conference on Learning Representations* (Vienna).

Aksan, N., Kısac, B., Aydın, M., and Demirbuken, S. (2009). Symbolic interaction theory. *Proc. Soc. Behav. Sci*. 1, 902–904. doi: 10.1016/j.sbspro.2009.01.160

Anderson, P. W. (1972). More is different. *Science* 177, 393–396. doi: 10.1126/science.177.4047.393

Atkeson, C. G., and Schaal, S. (1997). "Robot learning from demonstration," in *International Conference on Machine Learning*, ed D. H. Fisher Jr. (Nashville, TN: Morgan Kaufmann Publishers Inc.), 12–20.

Baillargeon, R. (2001). "Infants' physical knowledge: of acquired expectations and core principles," in *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler* (Cambridge, MA: The MIT Press), 341–361.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Barto, A. G., and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dyn. Syst. Theory Appl.* 13, 41–77. doi: 10.1023/A:1022140919877

Belsky, J., and Most, R. K. (1981). From exploration to play: a cross-sectional study of infant free play behavior. *Dev. Psychol.* 17, 630–639. doi: 10.1037/0012-1649.17.5.630

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., et al. (2020). "Experience grounds language," in *Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics), 8718–8735. doi: 10.18653/v1/2020.emnlp-main.703

Bordes, A., Lan Boureau, Y., and Weston, J. (2017). "Learning end-to-end goal-oriented dialog," in *International Conference on Learning Representations* (Toulon).

Botvinick, M., and Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philos. Trans. R. Soc. B Biol. Sci.* 369:1655. doi: 10.1098/rstb.2013.0480

Branavan, S. R. K., Zettlemoyer, L. S., and Barzilay, R. (2010). "Reading between the lines: learning to map high-level instructions to commands," in *Annual Meeting of the Association for Computational Linguistics, ACL '10* (Uppsala: Association for Computational Linguistics), 1268–1277.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, Vol. 33, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Vancouver, BC: Curran Associates, Inc.), 1877–1901.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2019). "Large-scale study of curiosity-driven learning," in *International Conference on Learning Representations*.

Burghardt, G. M. (2006). *The Genesis of Animal Play: Testing the Limits*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3229.001.0001

Chaplot, D. S., Sathyendra, K. M., Pasumarthi, R. K., Rajagopal, D., and Salakhutdinov, R. (2018). "Gated-attention architectures for task-oriented language grounding," in *Conference on Artificial Intelligence*, eds S. A. McIlraith and K. Q. Weinberger (New Orleans, LA: AAAI Press), 2819–2826.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., et al. (2019). "BabyAI: first steps towards grounded language learning with a human in the loop," in *International Conference on Learning Representations* (New Orleans, LA).

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford; New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780190217013.001.0001

Colas, C., Karch, T., Lair, N., Dussoux, J.-M., Moulin-Frier, C., Dominey, P., et al. (2020). "Language as a cognitive tool to imagine goals in curiosity driven exploration," in *Advances in Neural Information Processing Systems*, Vol. 33, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc.), 3761–3774.

Confalonieri, R., Eppe, M., Schorlemmer, M., Kutz, O., Pe naloza, R., and Plaza, E. (2018). Upward refinement operators for conceptual blending in the description logic EL ++. *Ann. Math. Artif. Intell.* 82, 69–99. doi: 10.1007/s10472-016-9524-8

Confalonieri, R., Eppe, M., Schorlemmer, M., Kutz, O., Peñaloza, R., and Plaza, E. (2015). "Upward refinement for conceptual blending in description logic "an ASP-based approach and case study in EL ++"," in *Workshop on Ontologies and Logic Programming for Query Answering*.

Confalonieri, R., Schorlemmer, M., Kutz, O., Peñaloza, R., Plaza, E., and Eppe, M. (2016). "*Conceptual blending in El++*," in *International Workshop on Description Logics* (Cape Town).

Côté, M.-A., Kádár, Á., Yuan, X., Kybartas, B., Barnes, T., Fine, E., et al. (2019). "Textworld: a learning environment for text-based games," in *Computer Games*, Vol. 1017, eds T. Cazenave, A. Saffidine, and N. Sturtevant (Springer International Publishing), 41–75. doi: 10.1007/978-3-030-24337-1_3

Cruz, F., Twiefel, J., Magg, S., Weber, C., and Wermter, S. (2015). "Interactive reinforcement learning through speech guidance in a domestic scenario," in *International Joint Conference on Neural Networks* (Killarney), 1–8. doi: 10.1109/IJCNN.2015.7280477

Dean, V., Tulsiani, S., and Gupta, A. (2020). "See, hear, explore: curiosity via audio-visual association," in *Advances in Neural Information Processing Systems*, Vol. 33, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc.), 14961–14972.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eppe, M., Gumbsch, C., Kerzel, M., Nguyen, P. D. H., Butz, M. V., and Wermter, S. (2020). Hierarchical principles of embodied reinforcement learning: a review.

Eppe, M., Maclean, E., Confalonieri, R., Kutz, O., Schorlemmer, M., Plaza, E., et al. (2018). A computational framework for conceptual blending. *Artif. Intell.* 256, 105–129. doi: 10.1016/j.artint.2017.11.005

Eppe, M., Nguyen, P. D. H., and Wermter, S. (2019). From semantics to execution: integrating action planning with reinforcement learning for robotic causal problem-solving. *Front. Robot. AI* 6:123. doi: 10.3389/frobt.2019.00123

Eppe, M., and Oudeyer, P.-Y. (2021). Intelligent behavior depends on the ecological niche: interview with Dr. Pierre–Yves Oudeyer. *Künstliche Intelligenz* 35, 103–108. doi: 10.1007/s13218-020-00696-1

Feldman, J., and Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain Lang.* 89, 385–392. doi: 10.1016/S0093-934X(03)00355-9

Feldman, J. A. (2006). *From Molecule to Metaphor: A Neural Theory of Language. A Bradford Book*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3135.001.0001

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850. doi: 10.1080/17470210701623605

Forestier, S., Mollard, Y., and Oudeyer, P.-Y. (2017). *Intrinsically motivated goal exploration processes with automatic curriculum* learning.

Frankland, S. M., and Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annu. Rev. Psychol.* 71, 273–303. doi: 10.1146/annurev-psych-122216-011829

Friederici, A. D., and Singer, W. (2015). Grounding language processing on basic neurophysiological principles. *Trends Cogn. Sci.* 19, 329–338. doi: 10.1016/j.tics.2015.03.012

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

Gamerschlag, T., Gerland, D., Osswald, R., and Petersen, W. (eds.). (2014). *Frames and Concept Types: Applications in Language and Philosophy, Volume 94 of Studies in Linguistics and Philosophy*. Springer International Publishing. doi: 10.1007/978-3-319-01541-5

Garagnani, M., and Pulvermüller, F. (2016). Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *Eur. J. Neurosci.* 43, 721–737. doi: 10.1111/ejn.13145

Georgie, Y. K., Schillaci, G., and Hafner, V. V. (2019). "An interdisciplinary overview of developmental indices and behavioral measures of the minimal self," in *International Conference on Development and Learning and Epigenetic Robotics* (Oslo: IEEE), 129–136. doi: 10.1109/DEVLRN.2019.8850703

Hafner, V. V., Loviken, P., Pico Villalpando, A., and Schillaci, G. (2020). Prerequisites for an artificial self. *Front. Neurorobot.* 14:5. doi: 10.3389/fnbot.2020.00005

Hahn, M., Krantz, J., Batra, D., Parikh, D., Rehg, J., Lee, S., et al. (2020). "Where are you? Localization from embodied dialog," in *Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics), 806–822. doi: 10.18653/v1/2020.emnlp-main.59

Haynes, J. D., Wisniewski, D., Gorgen, K., Momennejad, I., and Reverberi, C. (2015). "fMRI decoding of intentions: compositionality, hierarchy and prospective memory," *in International Winter Conference on Brain-Computer Interface* (Institute of Electrical and Electronics Engineers Inc.). doi: 10.1109/IWW-BCI.2015.7073031

Heinrich, S., Yao, Y., Hinz, T., Liu, Z., Hummel, T., Kerzel, M., et al. (2020). Crossmodal language grounding in an embodied neurocognitive model. *Front. Neurorobot*. 14:52. doi: 10.3389/fnbot.2020.00052

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., et al. (2017). Grounded language learning in a simulated 3D world. *arXiv preprint arXiv:1706.06551*.

Hill, F., Clark, S., Hermann, K. M., and Blunsom, P. (2019). Understanding early word learning in situated artificial agents. *arXiv preprint arXiv:1710.09867*.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., et al. (2020). "Environmental drivers of systematicity and generalization in a situated agent," in *International Conference on Learning Representations* (Addis Ababa).

Hill, F., Tieleman, O., von Glehn, T., Wong, N., Merzic, H., and Clark, S. (2021). "Grounded language learning fast and slow," in *International Conference on Learning Representations* (Vienna).

Hoffmann, M., Marques, H., Arieta, A., Sumioka, H., Lungarella, M., and Pfeifer, R. (2010). Body schema in robotics: a review. *IEEE Trans. Auton. Mental Dev*. 2, 304–324. doi: 10.1109/TAMD.2010.2086454

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

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., et al. (2019). "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, CA: PMLR), 3040–3049.

Jiang, Y., Gu, S. S., Murphy, K. P., and Finn, C. (2019). "Language as an abstraction for hierarchical deep reinforcement learning," in *Advances in Neural Information Processing Systems*, Vol. 32, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (Vancouver, BC: Curran Associates, Inc.), 9419–9431.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017). "CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning," *in Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1988–1997. doi: 10.1109/CVPR.2017.215

Kaplan, F. (2007). In search of the neural circuits of intrinsic motivation. *Front. Neurosci*. 1, 225–236. doi: 10.3389/neuro.01.1.1.017.2007

Kidd, C., and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron* 88, 449–460. doi: 10.1016/j.neuron.2015.09.010

Kiefer, M., and Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex* 48, 805–825. doi: 10.1016/j.cortex.2011.04.006

Lakoff, G., and Johnson, M. (2010). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York, NY: Basic Books.

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., et al. (2019). "A survey of reinforcement learning informed by natural language," in *International Joint Conference on Artificial Intelligence* (Macao), 6309–6317. doi: 10.24963/ijcai.2019/880

Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S., et al. (2020). "Learning latent plans from play," in *Conference on Robot Learning*, eds L. P. Kaelbling, D. Kragic, and K. Sugiura (PMLR), 1113–1132.

Lynch, C., and Sermanet, P. (2021). "Language Conditioned Imitation Learning Over Unstructured Data," in *Proceedings of Robotics: Science and Systems*. doi: 10.15607/RSS.2021.XVII.047

Madden, C., Hoen, M., and Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain Lang*. 112, 180–188. doi: 10.1016/j.bandl.2009.07.001

Madureira, B., and Schlangen, D. (2020). "An overview of natural language state representation for reinforcement learning," in *International Conference on Machine Learning*.

Mandler, J. M. (2004). Thought before language. *Trends Cogn. Sci*. 8, 508–513. doi: 10.1016/j.tics.2004.09.004

Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., and Fox, D. (2012). "A joint model of language and perception for grounded attribute learning," in *International Conference on Machine Learning* (Edinburgh: Omni Press), 1435–1442.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., and Schütze, H. (2020). Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*. doi: 10.5282/ubm/epub.72201

Mead, G. H., Morris, C. W., and Mead, G. H. (2000). *Mind, Self, and Society: From the Standpoint of a Social Behaviorist, Volume 1 of Works of George Herbert Mead*. Chicago: University of Chicago Press.

Mirolli, M., and Parisi, D. (2011). Towards a vygotskyan cognitive robotics: the role of language as a cognitive tool. *N. Ideas Psychol*. 29, 298–311. doi: 10.1016/j.newideapsych.2009.07.001

Misra, D., Langford, J., and Artzi, Y. (2017). "Mapping instructions and visual observations to actions with reinforcement learning," in *Conference on Empirical Methods in Natural Language Processing* (Copenhagen: Association for Computational Linguistics), 1004–1015. doi: 10.18653/v1/D17-1106

Narasimhan, K., Barzilay, R., and Jaakkola, T. (2018). Grounding language for transfer in deep reinforcement learning. *J. Artif. Intell. Res*. 63, 849–874. doi: 10.1613/jair.1.11263

Ng, A. Y., and Russell, S. J. (2000). "Algorithms for inverse reinforcement learning," in *International Conference on Machine Learning* (Stanford, CA: Morgan Kaufmann Publishers Inc.), 663–670.

Nguyen, P. D. H., Eppe, M., and Wermter, S. (2020). *Robotic self-representation improves manipulation skills and transfer* learning.

Nguyen, P. D. H., Georgie, Y. K., Kayhan, E., Eppe, M., Hafner, V. V., and Wermter, S. (2021). Sensorimotor representation learning for an "active self" in robots: a model survey. *Künstliche Intelligenz* 35, 9–35. doi: 10.1007/s13218-021-00703-z

Oh, J., Singh, S., Lee, H., and Kohli, P. (2017). "Zero-shot task generalization with multi-task deep reinforcement learning," in *International Conference on Machine Learning*, eds D. Precup and Y. W. Teh (Sydney, NSW: PMLR), 2661–2670.

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

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychol. Rev*. 76, 241–263. doi: 10.1037/h0027272

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). "Curiosity-driven exploration by self-supervised prediction," in *International Conference on Machine Learning* (Sydney, NSW), 2778–2787. doi: 10.1109/CVPRW.2017.70

Paul, R., Arkin, J., Aksaray, D., Roy, N., and Howard, T. M. (2018). Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *Int. J. Rob. Res*. 37, 1269–1299. doi: 10.1177/0278364918777627

Paulus, M. (2014). How and why do infants imitate? An ideomotor approach to social and imitative learning in infancy (and beyond). *Psychon. Bull. Rev*. 21, 1139–1156. doi: 10.3758/s13423-014-0598-1

Philippsen, A. (2021). Goal-directed exploration for learning vowels and syllables: a computational model of speech acquisition. *Künstliche Intelligenz* 35, 53–70. doi: 10.1007/s13218-021-00704-y

Piaget, J. (1926). *The Language and Thought of the Child*. Brace: Harcourt.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., et al. (2021). Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci*. 21, 188–194. doi: 10.1016/S0166-2236(98)01260-0

Röder, F., Eppe, M., Nguyen, P. D. H., and Wermter, S. (2020). "Curious hierarchical actor-critic reinforcement learning," *in International Conference on Artificial Neural Networks* (Bratislava), 408–419. doi: 10.1007/978-3-030-61616-8_33

Saleh, A., Jaques, N., Ghandeharioun, A., Shen, J., and Picard, R. (2020). Hierarchical reinforcement learning for open-domain dialog. *Conf. Artif. Intell*. 34, 8741–8748. doi: 10.1609/aaai.v34i05.6400

Schillaci, G., Hafner, V. V., and Lara, B. (2016). Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents. *Front. Robot. AI* 3:39. doi: 10.3389/frobt.2016.00039

Schillaci, G., Hafner, V. V., Lara, B., and Grosjean, M. (2013). "Is that me? Sensorimotor learning and self-other distinction in robotics," in

International Conference on Human-Robot Interaction (IEEE), 223–224. doi: 10.1109/HRI.2013.6483582

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., and Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife* 8:e41703. doi: 10.7554/eLife.41703

Shastri, L. (1999). Recruitment of binding and binding-error detector circuits via long-term potentiation. *Neurocomputing* 26–27, 865–874. doi: 10.1016/S0925-2312(98)00131-3

Shi, W., and Yu, Z. (2018). "Sentiment adaptive end-to-end dialog systems," in *Annual Meeting of the Association for Computational Linguistics* (Melbourne, VIC: Association for Computational Linguistics), 1509–1519. doi: 10.18653/v1/P18-1140

Shin, Y. K., Proctor, R. W., and Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychol. Bull.* 136, 943–974. doi: 10.1037/a0020541

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., et al. (2020). "ALFRED: a benchmark for interpreting grounded instructions for everyday tasks," in *Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE). doi: 10.1109/CVPR42600.2020.01075

Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., and Hausknecht, M. (2021). "ALFWorld: aligning text and embodied environments for interactive learning," in *International Conference on Learning Representations*.

Sohn, K., Lee, H., and Yan, X. (2015). "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, Vol. 28, eds C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Montréal, QC: Curran Associates, Inc.).

Spranger, M., Suchan, J., Bhatt, M., and Eppe, M. (2014). "Grounding dynamic spatial relations for embodied (robot) interaction," in *Pacific Rim International Conferences on Artificial Intelligence*, 958–971. doi: 10.1007/978-3-319-13560-1_83

Steels, L. (2007). *The Symbol Grounding Problem Has Been Solved. So What's Next? Symbols, Embodiment and Meaning*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199217274.003.0012

Steels, L., and Loetzsch, M. (2012). "The grounded naming game," in *Advances in Interaction Studies*, Vol. 3, ed L. Steels (Amsterdam: John Benjamins Publishing Company), 41–59. doi: 10.1075/ais.3.04ste

Steels, L., Spranger, M., Trijp, R. V., Höfer, S., and Hild, M. (2012). "Emergent action language on real robots," in *Language Grounding in Robots*, eds L. Steels and M. Hild (Boston, MA: Springer), 255–276. doi: 10.1007/978-1-4614-3064-3_13

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton-Smith, B. (2001). *The Ambiguity of Play*. Cambridge, MA: Harvard University Press.

Tani, J. (2016). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. Oxford University Press. doi: 10.1093/acprof:oso/9780190281069.001.0001

Tellex, S., Gopalan, N., Kress-Gazit, H., and Matuszek, C. (2020). Robots that use language. *Annu. Rev. Control Robot. Auton. Syst.* 3, 25–55. doi: 10.1146/annurev-control-101119-071628

Trott, S., Eppe, M., and Feldman, J. (2016). "Recognizing intention from natural language: clarification dialog and construction grammar," in *Workshop on Communicating Intentions in Human-Robot Interaction, International Symposium on Human and Robot Interactive Communication* (New York, NY: IEEE).

Turner, M. (2015). *The Origin of Ideas: Blending, Creativity, and the Human Spark*. (Oxford University Press).

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., and Wermter, S. (2021). Survey on reinforcement learning for language processing. *arXiv preprint arXiv:2104.05565*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vygotsky, L. (1967). Play and its role in the mental development of the child. *J. Russ. East Eur. Psychol.* 5, 6–18. doi: 10.2753/RPO1061-040505036

Waxman, S. R., and Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cogn. Psychol.* 29, 257–302. doi: 10.1006/cogp.1995.1016

Wermter, S., Page, M., Knowles, M., Gallese, V., Pulvermüller, F., and Taylor, J. (2009). Multimodal communication in animals, humans and robots: an introduction to perspectives in brain-inspired informatics. *Neural Netw.* 22, 111–115. doi: 10.1016/j.neunet.2009.01.004

Willems, R. M., Hagoort, P., and Casasanto, D. (2010). Body-specific representations of action verbs: neural evidence from right- and left-handers. *Psychol. Sci.* 21, 67–74. doi: 10.1177/0956797609354072

Wood, D., Bruner, J. S., and Ross, G. (1976). The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* 17, 89–100. doi: 10.1111/j.1469-7610.1976.tb00381.x

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., et al. (2020). "Multilingual universal sentence encoder for semantic retrieval," in *Annual Meeting of the Association for Computational Linguistics, System Demonstrations* (Association for Computational Linguistics), 87–94. doi: 10.18653/v1/2020.acl-demos.12

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychol. Bull.* 133, 273–293. doi: 10.1037/0033-2909.133.2.273

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

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

frontiers
in Psychology

REVIEW
published: 16 September 2021
doi: 10.3389/fpsyg.2021.716220

The Interplay Between Affective Processing and Sense of Agency During Action Regulation: A Review

Jakob Kaiser, Madalina Buciuman, Sandra Gigl, Antje Gentsch and Simone Schütz-Bosbach*

LMU Munich, Department of Psychology, General and Experimental Psychology, Munich, Germany

Sense of agency is the feeling of being in control of one's actions and their perceivable effects. Most previous research identified cognitive or sensory determinants of agency experience. However, it has been proposed that sense of agency is also bound to the processing of affective information. For example, during goal-directed actions or instrumental learning we often rely on positive feedback (e.g., rewards) or negative feedback (e.g., error messages) to determine our level of control over the current task. Nevertheless, we still lack a scientific model which adequately explains the relation between affective processing and sense of agency. In this article, we review current empirical findings on how affective information modulates agency experience, and, conversely, how sense of agency changes the processing of affective action outcomes. Furthermore, we discuss in how far agency-related changes in affective processing might influence the ability to enact cognitive control and action regulation during goal-directed behavior. A preliminary model is presented for describing the interplay between sense of agency, affective processing, and action regulation. We propose that affective processing could play a role in mediating the influence between subjective sense of agency and the objective ability to regulate one's behavior. Thus, determining the interrelation between affective processing and sense of agency will help us to understand the potential mechanistic basis of agency experience, as well as its functional significance for goal-directed behavior.

Keywords: sense of agency, emotions, cognitive control, feedback processing, action regulation

OPEN ACCESS

Edited by:
Stephan Alexander Verschoor,
Leiden University, Netherlands

Reviewed by:
Wilfried Kunde,
Julius Maximilian University of Würzburg, Germany
Pablo Lanillos,
Radboud University Nijmegen, Netherlands

****Correspondence:***
Jakob Kaiser
j.kaiser@psy.lmu.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *28 May 2021*
Accepted: *20 August 2021*
Published: *16 September 2021*

Citation:
Kaiser J, Buciuman M, Gigl S, Gentsch A and Schütz-Bosbach S (2021) The Interplay Between Affective Processing and Sense of Agency During Action Regulation: A Review. Front. Psychol. 12:716220. doi: 10.3389/fpsyg.2021.716220

INTRODUCTION

To effectively reach our goals, it is important to assess how much influence we have over our environment. Sense of agency is the subjective feeling of being in control of one's actions and their perceivable effects (Gentsch and Schütz-Bosbach, 2015; Haggard, 2017). An inflated sense of agency has been associated with irrational and potentially self-destructive actions. For example, gambling addicts can have an unrealistically high feeling of control over chance outcomes (Orgaz et al., 2013). A diminished sense of agency has been related to inaction and a lack of perseverance (Bhanji et al., 2016; Studer et al., 2020). Strong feelings of loss of control have been associated with depression and anxiety disorders (Gallagher et al., 2014; Maier and Seligman, 2016). Thus, it is important to determine how sense of agency is established and through which mechanisms it influences our behavior.

A fundamental goal of instrumental actions is to receive positive outcomes and to avoid negative consequences. Therefore, we appear to constantly monitor the affective value of action outcomes. Since affective feedback is crucial for self-determined actions, it has been proposed that our sense of agency is bound to the processing of affective information (Synofzik et al., 2013; Gentsch and Synofzik, 2014; Ly et al., 2019). However, most previous research so far focussed on non-affective, sensory, and cognitive determinants of sense of agency. As a consequence, the potential functional relevance of affective processing for agency experience is not yet clearly understood.

This article aims to give an overview of current research concerning the interplay between affective processing and sense of agency. More specifically, we will discuss the empirical evidence regarding two questions (1) Do affective information or emotional states exert an influence on sense of agency? (2) Does sense of agency influence the processing of affective information and, particularly, on how humans process affective feedback during goal-directed actions? To preview our conclusions, current findings provide evidence for a bidirectional relation between affective processing and sense of agency. However, many details of the potential interaction between affect and agency experience still need to be clarified.

In the last part of this article, we will discuss the potential practical implications of the link between sense of agency and affective processing. This discussion will be guided by a tentative model about the interrelation between sense of agency, affective processing, and action regulation. For the purpose of this review, we define action regulation as the goal-oriented adjustment of ongoing or habitual behavior in response to environmental demands. In a nutshell, we propose that affective processing could play a role in mediating the influence between subjective sense of agency and the objective ability to regulate one's behavior. While an enhanced sense of agency might facilitate action regulation by heighten one's sensitivity toward task-relevant affective feedback, a diminished sense of agency in contrast could lead to blunted processing of affective feedback, resulting in less effective behavioral regulation. While more empirical findings are needed to critically evaluate this model, investigating the relationship between sense of agency and affective processing could help to elucidate the role of sense of agency for goal-directed behavior.

DETERMINANTS OF SENSE OF AGENCY

A number of different psychological terms are used in the literature to describe the subjective feeling of having or lacking agency over one's actions and the environment, such as sense of agency, self-efficacy, control beliefs, illusion of control, or learned helplessness (Ly et al., 2019). Control beliefs have sometimes been assessed as personality traits, meaning that people can maintain relative stable assumptions about their general degree of control over the environment (Craig et al., 1984; Galvin et al., 2018). In contrast, sense of agency is commonly meant to describe a psychological state, which can potentially fluctuate over time (Moore, 2016). For example, having success in learning a new skill can lead to a gradual increase in sense of agency (van der Wel et al., 2012). Accordingly, this article will focus on studies which manipulate or measure changes in agency experience in an experimental setting. For investigations of the relation between trait control beliefs and affect see for example: Gallagher et al. (2014), Harnett et al. (2015), and Koffer et al. (2019).

Previous research has identified many perceptual and cognitive factors which can increase subjective sense of agency. For example, participants are more likely to assume agency over sensory effects in the environment, if these effects follow their own actions in a predictable way and in close temporal proximity (Haggard and Tsakiris, 2009; Gentsch and Schütz-Bosbach, 2015). Thus, in many circumstances, sense of agency for our action is based on perceptual and cognitive processes.

The results of our actions often have personal relevance. Being able to produce perceivable sensory effects with our own actions can feel inherently pleasurable or motivating (Eitam et al., 2013; Karsh and Eitam, 2015). Moreover, we often engage in actions because we believe they might lead to rewarding or pleasant consequences, or they might help us to avoid punishment or detrimental outcomes. For example, during reinforcement learning, positive or negative feedback is usually provided after each action to either reinforce or discourage our current behavior (Cockburn et al., 2014; Kaiser et al., 2021). These types of action outcomes are not merely sensory events, but they can evoke affective states. One common approach to classify affective stimuli or states is to distinguish between positive and negative valence (Russell, 2003; Posner et al., 2005). In the context of goal-directed actions, action outcomes with positive valence occur when our actions lead to events which we perceive as pleasant or desirable, such as reward or praise. Action outcomes with negative valence are action effects which participants perceive as unpleasant or aversive, such as error messages or monetary losses. In the following review, we will summarize in how far the positive or negative valence influences sense of agency and, conversely, how sense of agency can influence the affective processing and affective states, such as the positive or negative feelings of participants.

DOES AFFECT INFLUENCE SENSE OF AGENCY?

This section will summarize experiments which tested the influence of affective context or emotional stimuli on sense of agency. More specifically, the guiding question is: Does positive compared to negative affect lead to an increase or decrease in sense of agency? Relevant studies manipulated affect-related aspects of an experimental task while measuring participants' sense of agency. Two types of affective manipulations were commonly used. Most studies manipulated the affective value of action effects, for example by letting participants perform an action which either led to the appearance of positive or negative action outcomes. This allowed to measure participants' sense of agency over positive compared to negative action effects. Some studies manipulated the affective context of an otherwise neutral action-effect sequence, for example via mood induction directly prior to an action. This allowed to test if participants' feelings can bias their sense of agency, even when affect is incidental to the action and its effect in question.

For measuring sense of agency, studies either employed explicit or implicit approaches. Explicit measures rely on self-report of agency experience, for example by asking people to rate their own perceived feeling of control over the outcome of each trial. It has been suggested that explicit self-report of agency might not always be accurate, e.g., because of demand effects or hindsight biases (Synofzik et al., 2008; Haggard, 2017). Therefore, some studies rely on implicit measures which avoid self-report. The most common implicit measures of sense of agency are sensory attenuation and temporal binding. Sensory attenuation describes the phenomenon that self-produced compared to passively perceived sensory effects lead to lower perceptual and neural impact (Blakemore et al., 2000; Gentsch et al., 2012). Temporal binding is a perceptual bias in which the delay between an action and the ensuing effect (e.g., a button press and a subsequent sound) is perceived to be shorter in time when the action is performed by oneself than by someone else (Haggard et al., 2002; Wolpe and Rowe, 2014). Based on these phenomena, many studies assume that stronger sensory attenuation or temporal binding indicate an increase in sense of agency. Since studies employing either explicit or implicit measures found partially divergent results, we will discuss relevant findings separately for explicit measures (i.e., self-report) and implicit measures (intentional binding or sensory attenuation).

Influence of Affective Manipulations on Self-Report of Agency

Most studies using self-report measures found that positive compared to negative action outcomes lead to increased sense of agency. This has been shown for different types of affective stimuli such as consonant/dissonant sounds (Barlas and Obhi, 2014; Barlas et al., 2017), emotional facial expressions (Gentsch et al., 2015), and performance feedback in gambling tasks (Kulakova et al., 2017; Herman and Tsakiris, 2020) or motor control tasks (Oishi et al., 2018, 2019; Le Bars et al., 2020). The finding of increased sense of agency for positive action effects has been interpreted as part of a self-serving bias in human cognition (Gentsch and Synofzik, 2014; Haggard, 2017). Humans are more likely to attribute positive than negative events toward themselves (Mezulis et al., 2004). Thus, positive outcomes are more likely to be associated with increased sense of agency.

The studies described so far manipulated the affective valence of action effects by presenting either positive or negative action outcomes. Another approach to investigate the influence of affect on sense of agency would be to directly manipulate participants' current mood states. However, there is little evidence that mood manipulations influence explicit sense of agency. One study found that the induction of stress via the Trier Social Stress Test had no effect on agency ratings for otherwise neutral action effects (Stern et al., 2020). Since stress is usually experienced as a strongly negative affective state, this suggests that participants do not necessarily integrate their current feeling state in explicit agency judgements. More studies are needed to clarify if the affective context of an action might bias agency experience for unrelated, neutral action outcomes.

Influence of Affective Manipulations on Implicit Agency Measures

Several studies tested if the valence of action outcomes has an effect on implicit measures of agency. For intentional binding, the evidence for affect-specific influences is mixed. Some studies found that negative compared to neutral or positive action outcomes decrease temporal binding (Takahata et al., 2012; Yoshie and Haggard, 2013; Barlas and Obhi, 2014; Borhani et al., 2017; Haggard, 2017; Nataraj et al., 2020). Since less temporal binding is assumed to indicate lower sense of agency, these findings are consistent with studies employing explicit agency measures, which found that negative action outcomes were associated with lower sense of agency.

However, there are also a number of studies which did not find any effect of outcome valence on temporal binding (Barlas et al., 2017, 2018; Kulakova et al., 2017; Moreton et al., 2017; Herman and Tsakiris, 2020). The absence of valence-specific binding effects in some studies might indicate that the valence of action outcomes only influences temporal binding under specific circumstances. In line with this assumption, a few experiments found that the effects of outcome valence on temporal binding depend on the predictability of action effects. Some studies reported that positive compared to negative outcomes only led to stronger temporal binding when the task context allowed to reliably predict if an action would lead to a positive or negative effect (Yoshie and Haggard, 2017). In contrast, when the valence of action outcomes was not predictable, no valence-specific binding effects were found. However, other studies found the opposite pattern of results, with increased binding for positive effects only for unpredictable, but not for predictable, outcomes (Christensen et al., 2016; Tanaka et al., 2020). Lastly, one study reported increased binding for predictable compared to unpredictable electric shocks (meaning strongly negative stimulation) as effects of one's own actions (Beck et al., 2017). Overall, these studies might indicate that the impact of outcome valence on agency interacts with other factors, such as anticipation and stimulus predictability. However, the exact nature of this interaction is not consistent across studies, and therefore not clearly understood.

Some experiments found that the valence of action effects might not only influence temporal binding between the action and the ensuing effect itself but could also have an impact on subsequent actions in the same task. For reinforcement learning tasks, it was found that negative compared to positive performance feedback on a trial increased intentional binding for actions on the subsequent trial (Di Costa et al., 2018; Majchrowicz et al., 2020). Errors are known to evoke increased top-down control of one's behavior to improve performance on subsequent trials (Ullsperger et al., 2014). Therefore, stronger binding after errors could indicate that engaging in top-down control is related to an enhanced sense of agency (Majchrowicz et al., 2020).

Only very few studies measured the effect of affective valence on sensory attenuation with mixed results. Some found stronger sensory attenuation for positive compared to negative action outcomes (Gentsch et al., 2015), others reported evidence for stronger attenuation for more negative action effects (Borhani et al., 2017; Osumi et al., 2019; Majchrowicz and Wierzchoń, 2021), or no effect of outcome valence on attenuation (Beck et al., 2017). Thus, it is currently not clear under which circumstances the valence of action outcomes modulates sensory attenuation.

As for explicit measures, there are less studies about the influence of participants' mood state on implicit measures of sense of agency for neutral action-effect sequences. Some studies reported that positive mood inductions prior to actions can increase temporal binding, while negative mood inductions led to decreased binding effects (Aarts et al., 2012; Obhi et al., 2013; Christensen et al., 2019). This could be seen as evidence suggesting that participants' affective state can bias their feeling of agency on an implicit level, with positive compared to negative mood increasing sense of agency.

Summary: Influence of Affect on Sense of Agency

To summarize, several studies measured the effect of positive or negative action outcomes on sense of agency. Experiments relying on self-report show a mostly consistent pattern: Positive compared to negative action outcomes increase the explicit feeling of agency. For studies employing implicit measures the results are more varied and partly contradictory. There is evidence that positive compared to negative action outcomes either increase, decrease, or do not influence implicit sense of agency. At the very least, this indicates the need to identify additional factors which determine the impact of affective information on temporal binding and sensory attenuation. Importantly, there is evidence that sensory attenuation and temporal binding can be influenced by other factors than personal agency, such as the temporal predictability of action effects or changes in attention (Buehner and May, 2003; Kok et al., 2012; Kaiser and Schütz-Bosbach, 2018). Thus, it is not clear in how far the divergent results found via sensory attenuation or temporal binding capture genuine differences in agency experience, rather than confounding factors specific to the implicit measures itself.

There are very few reports about the influence of affective context, such as participants' mood, on sense of agency for neutral action effects. Some studies, mostly relying on temporal binding, suggest that positive compared to negative mood might increase sense of agency for unrelated action effects. It remains to be seen if similar effects can be found for explicit measures of agency. Moreover, future studies could consider the possibility that the impact of affective states on sense of agency depends on interindividual differences in affective processing. For example, individuals with diminished emotional coping skills might be more likely to infer agency from their current feelings.

DOES SENSE OF AGENCY INFLUENCE AFFECTIVE PROCESSING?

The following section will discuss experiments about the influence of agency experience on affective processing. Several approaches exist for the experimental manipulation of sense of agency (cf. **Box 1**). Most studies concerning the influence of agency experience on affect manipulated agency by varying the degree of choice (choice agency) or the degree of outcome reliability (outcome agency). For example, many studies compared the impact of rewards or losses which were either the result of forced-choice or free-choice actions. Such experiments allow measuring the effect of high compared to low sense of agency on affect-related measures. Two types of measures can be distinguished. First, some studies investigated the effect of agency manipulations on participants' affective state, for example by testing if changes in sense of agency influenced participants' mood. Second, other studies investigated the effect of agency manipulations on participants' sensitivity for stimuli with positive or negative valence. This allowed to test whether high compared to low sense of agency increased the subjective or neural impact of affective feedback. Answering this question would help to clarify if sense of agency influences the way we process affective information, such as positive or negative feedback during performance tasks. We will first discuss studies investigating agency effects on participants' affective states, and subsequently summarize experiments dealing with the influence of agency on the sensitivity for affective feedback.

Influence of Sense of Agency on Affective States

Several studies tested if sense of agency influences participants' self-reported emotional states. Most experiments found that having a degree of choice over one's actions and/or a feeling of control over ensuing action effects led to more positive or less negative affect (Abelson et al., 2008; Thuillard and Dan-Glauser, 2017, 2020; Stolz et al., 2020; Li et al., 2021). Moreover, participants prefer tasks which allow them to make choices compared to tasks where they cannot choose between different options, even when their own choices are not more likely to result in better outcomes (Leotti and Delgado, 2011, 2014; Fujiwara et al., 2013; Cockburn et al., 2014; Mistry and Liljeholm, 2016; Bobadilla-Suarez et al., 2017; Wang and Delgado, 2019). Items which are obtained through one's own choice are subjectively judged as being more valuable (Fujiwara et al., 2013). On a neural level, the mere anticipation of being able to make a choice has been found to increase activity in brain regions which are linked to reward processing, such as the ventral striatum (Tricomi et al., 2004; Bjork and Hommer, 2007; Leotti and Delgado, 2014; Lorenz et al., 2015; Romaniuk et al., 2019; Wang and Delgado, 2019; Stolz et al., 2020). Overall, these studies suggest that increased sense of agency is commonly experienced as desirable, and leads to increased positive affect (Leotti et al., 2010).

While having some degree of choice can increase positive affect, being presented with too many options can lead to increased negative, not positive, feelings (Iyengar and Lepper, 2000; Reutskaja and Hogarth, 2009). Having to consider a high number of different options might lead to information overload and, thus, higher cognitive demand (Scheibehenne et al., 2010; Chernev et al., 2012). Thus, the positive effects of choice agency can potentially be diminished or even be reversed in contexts where increased freedom of choice significantly increases task difficulty (Greifeneder et al., 2010).

BOX 1 | Experimental manipulations of sense of agency.

Different techniques: Experimental manipulations of agency typically aim at inducing a high or low sense of agency in participants to investigate the effect of agency experience on other psychological measures of interest. Techniques to manipulate agency can target different aspects of goal-directed behavior, and therefore differ widely across studies. At least three different types of manipulations can be distinguished:

- **Motor agency**: Many studies investigate agency at the level of motor executions, for example by comparing a condition where participants actively elicit a motor action to produce a sensory effect (high motor agency), with a condition where they just passively perceive the same effect (low motor agency; e.g., Baess et al., 2011; Kaiser and Schütz-Bosbach, 2018). Thus, agency in this case means to trigger an outcome with one's own motor action.

- **Choice agency**: Some studies manipulate the degree of choice over what type of action participants perform, for example by comparing a condition where participants can choose one of several buttons to press (free choice), with a condition where they have to press a predetermined button (forced choice; e.g., Fujiwara et al., 2013; Chambon et al., 2020). Agency in this case means to be able to choose between different actions with potentially different outcomes.

- **Outcome agency**: Since our actions are usually aimed at producing specific effects, such as obtaining rewards, we are more likely to feel in control when we can reliably produce the desired outcome (Moscarello and Hartley, 2017; Ly et al., 2019). Accordingly, some experiments manipulate agency experience by ensuring either that it is possible to produce a positive outcome (e.g., via highly reliable action-effect contingencies; high outcome agency), or giving participants no reliable chance to achieve the desired outcome (e.g., via random action-effect contingencies, low outcome agency; e.g., Nataraj et al., 2020; Li et al., 2021). Agency here means the ability to influence the environment in a way which is desirable to the agent.

Real vs. illusionary agency: Sense of agency is a subjective state, which can deviate from our objective level of control. Thus, sense of agency can be induced via real or illusionary agency. Inducing real agency means to provide an actual degree of control, for example by providing meaningful choices in a task. Inducing imaginary agency means to create an illusion of control, for example by making participants believe that outcomes in a task are dependent on their actions when in fact they are predetermined by the experimenter (e.g., Lorenz et al., 2015; Mühlberger et al., 2017). While providing an actual degree of control can lead to a more realistic task setting, inducing only the illusion of control might allow to more clearly attribute any experimental effect to changes in participants' subjective sense of agency, rather than other effects related to their objective mastery over the task.

Do different agency manipulations target the same processes? In many practical tasks, different aspects of agency are confounded. Importantly, it is unclear in how far different types of agency manipulations target the same or different cognitive and neural mechanisms. For example, a recent study reported that, compared to a condition where participants passively received rewarding outcomes (no agency), the neural processing of rewards was enhanced when participants performed a freely chosen action which triggered the rewarding outcome (motor and choice agency), but not when they had to perform a predetermined action to obtain the same reward (motor agency only; Hassall et al., 2019). This suggests that choice agency compared to motor agency might have different effects on the neural processing of action outcomes. More research is needed to clarify the potential differentiation between sense of agency on the level of motor execution (motor agency), action selection (choice agency), or outcome contingencies (outcome agency).

Several studies investigated the influence of sense of agency on neural or subjective measures of pain. Most of these experiments provided participants with some (real or illusionary) possibility to control the presence or duration of painful stimulation. Compared to a condition where participants experienced the same degree of pain stimulation without any form of control, the feeling of having agency usually led to lower self-reported levels of pain intensity, as well as less activity in brain areas associated with pain processing (Salomons et al., 2004, 2014; Wiech et al., 2006; Vancleef and Peters, 2011; Mohr et al., 2012; Szczepanowski et al., 2013; Bräscher et al., 2016). While pain is usually not considered to be an affective state, it is commonly associated with strong negative affect. Therefore, these findings are consistent with the notion that increased sense of agency can lower negative affect.

To conclude, most studies indicate that heightened sense of agency increases positive and/or decreases negative affect. However, an overabundance of choice might lead to aversive affective reactions in contexts where the decision-making process strongly increases task demand.

Influence of Sense of Agency on the Processing of Affective Stimuli

Several studies investigated if sense of agency increases or decreases the sensitivity for affective stimuli. Most experiments concerned with this question manipulated participants' sense of agency for positive or negative task feedback during learning or gambling tasks, while measuring neural correlates of feedback sensitivity via EEG. Commonly used measures entailed ERPs like the reward positivity component, a midcentral positive deflection which tends to be increased for positive compared to negative feedback (Proudfit, 2015). This component is also often reported as error negativity, which is calculated as the difference in reward positivity between negative and positive stimuli (Mühlberger et al., 2017). Other studies measured the P300 or oscillatory midfrontal theta power, both of which tend to show increased activity during task-relevant expectation violations and errors (Polich, 2007; Kaiser et al., 2019).

Most studies reported that high compared to low agency increased the neural responses for affective action outcomes. This has been found for the reward positivity/error negativity component (Yeung et al., 2005; Bellebaum et al., 2010; Li et al., 2011; Martin and Potts, 2011; Bismark et al., 2013; Legault and Inzlicht, 2013; Bellebaum and Colosio, 2014; Meng and Ma, 2015; Mühlberger et al., 2017; Mei et al., 2018; Yi et al., 2018; Hassall et al., 2019; Fang et al., 2020; Zheng et al., 2020), as well as for the P300 (Bellebaum et al., 2010; Mühlberger et al., 2017; Mei et al., 2018; Yi et al., 2018; Hassall et al., 2019; Fang et al., 2020), and midfrontal theta power (Zheng et al., 2020). Overall, these findings suggest that sense of agency increases the neural impact of affective feedback.

Studies reporting that sense of agency increases the neural impact for affective feedback might appear to be inconsistent with the phenomenon of sensory attenuation. As discussed above, sensory attenuation refers to the finding that sense of agency leads to lower, not higher, neural impact for self-produced action effects (Baess et al., 2011; Gentsch and Schütz-Bosbach, 2015). Importantly, sensory attenuation has most often been reported for non-affective action effects with little or no direct relevance for participants. In contrast, affective stimuli often have practical significance for humans. For example, positive or negative action outcomes can provide feedback over our current performance during goal-directed tasks. Thus, sense of agency might increase the impact of affective and task-relevant, but not of non-affective incidental action effects, to highlight the most self-relevant results of our own actions.

Moreover, studies investigating agency effect for non-affective vs. affective stimuli tend to differ with respect to the type of agency manipulation (cf. **Box 1**): Sensory attenuation for non-affective stimuli has been mostly found when manipulating motor agency, usually by comparing passive perception with active production of sensory effects (Blakemore et al., 1998; Weiss et al., 2011). In contrast, neural enhancement for affective stimuli has most often been reported for studies which manipulated choice and/or outcome agency, for example by comparing free-choice with forced-choice tasks (Li et al., 2011; Mühlberger et al., 2017; Mei et al., 2018). Accordingly, the occurrence of neural attenuation compared to neural enhancement might partly be related to which type of agency (i.e., motor/choice/outcome) is being manipulated (Hassall et al., 2019).

Lastly, sensory attenuation was commonly assessed via early markers of sensory processing, such as the N100 component in EEG (Baess et al., 2011; Timm et al., 2016). In contrast, neural enhancement for affective stimuli was usually found for frontocentral indicators of reward and punishment processing, such as the midfrontal reward positivity or P300. Thus, we cannot exclude the possibility that sense of agency is more likely to lead to an attenuation of early neural markers of sensory impact, but an enhancement of neural activity related to evaluative processing.

Overall, most current studies show that sense of agency can increase the neural impact of affective stimuli. We still lack sufficient empirical data to fully explain the divergent findings between agency effects for non-affective vs. affective action effects. It will be important to determine under which circumstances increased sense of agency leads to neural attenuation compared to neural enhancement, for example by investigating the role of task-relevance (task-relevant vs. incidental action effects), type of agency experience (via independent manipulations of motor/choice/outcome agency), and the neural processing stage (by comparing effect on neural components related to early sensory vs. evaluative processing).

Does Sense of Agency Lead to a Valence-Specific Bias in Neural Processing?

While many studies show that higher sense of agency increases the neural impact of affective feedback, it is less clear if these agency-related effects on affective processing are equally strong for positive and negative stimuli. Determining if agency experience leads to a selective enhancement of either positive or negative feedback is important, because such a finding would imply that sense of agency generates a valence-specific processing bias. One study found behavioral evidence for an agency-related positivity bias in a reinforcement learning task. High compared to low sense of agency led to selective increases in learning rates after positive, but not negative, feedback (Chambon et al., 2020). Such a selective enhancement of positive feedback could help to explain self-serving biases in the evaluation of one's own actions (Mezulis et al., 2004).

On a neural level, the evidence that sense of agency induces a valence-specific bias is less conclusive. Many studies do not test for potential valence-specific effects of sense of agency. Some of the experiments which address this question find that agency affects the neural processing of both positive and negative feedback to a similar degree (Mei et al., 2018; Hassall et al., 2019). However, others report that increased sense of agency more strongly enhances the neural impact of negative feedback (Bellebaum et al., 2010; Martin and Potts, 2011; Legault and Inzlicht, 2013), or positive feedback (Mühlberger et al., 2017). The inconsistency between studies might in part reflect the fact that studies concentrate on neural measures which are either more strongly related to reward sensitivity (midfrontal reward positivity) or the processing of errors and expectation violations (P300/midfrontal theta power). In line with this assumption, one study found that increased sense of agency led to an increased reward positivity component in the EEG for positive (but not negative) feedback, but increased midfrontal theta power for negative (but not positive) feedback (Zheng et al., 2020). This suggests that sense of agency increases the neural impact of both positive and negative feedback, albeit for different neural processes.

Summary: Influence of Sense of Agency on Affect

To summarize, research indicates that heightened sense of agency increases positive affect. However, while free choice over some task-relevant aspects can be employed to induce increased sense of agency, an overabundance of choice options might intensify task complexity and thus lead to negative affect. Furthermore, while heightened sense of agency is commonly assumed to lead to lower neural impact for non-affective stimuli, it has been found to lead to increased neural impact of affective feedback. Further research is needed to determine if agency-related changes in neural processing of affective feedback occur for both positive and negative feedback to the same degree, or if agency induces a valence-specific bias in neural processing, in the sense of a selective increase in sensitivity for either positive or negative feedback.

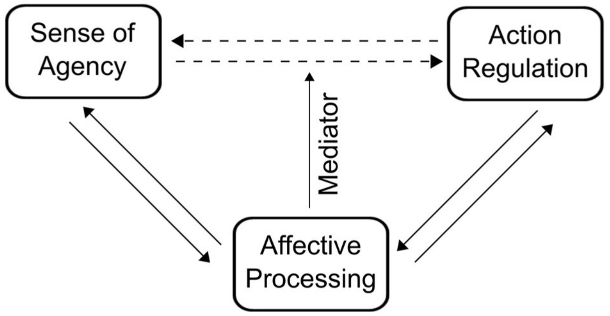


FIGURE 1 | Schematic representation of the proposed relation between sense of agency, affective processing, and action regulation. The model assumes a bidirectional relation between affective processing and sense of agency, as well as between affective processing and action regulation. Importantly, affective processing partially mediates the influence between sense of agency and action regulation.

THE ROLE OF SENSE OF AGENCY AND AFFECTIVE PROCESSING FOR ACTION REGULATION

The previous sections have shown that current research provides evidence for a bidirectional relationship between sense of agency and affect-related processes. Changes in affective states are associated with changes in sense of agency, and changes in sense of agency can alter affective states and the processing of affective stimuli, such as positive or negative performance feedback. This section will discuss the potential functional implications of the interaction between affective processing and sense of agency. This discussion focusses on a tentative model of the relationship between sense of agency, affect, and action regulation (**Figure 1**).

Action regulation in this context refers to an adjustment of ongoing behavior in order to improve one's chances to successfully reach a goal. Successful action regulation commonly depends on voluntary exertion of cognitive control mechanisms to override goal-incompatible behavioral tendencies (van de Vijver et al., 2011; Gratton et al., 2017; Kaiser and Schütz-Bosbach, 2019, 2021). Concerning the role of affective processing, we specifically focus here on the processing of positive and negative performance feedback during tasks which necessitate action regulation. We suggest that the interaction between sense of agency and affective processing plays a role in this process, since changes in sense of agency can either increase or dampen the sensitivity for affective task feedback (Bhanji and Delgado, 2014; Mühlberger et al., 2017; Hassall et al., 2019). Since behavioral adaption relies on the accurate processing of affective feedback, agency-related changes in affective processing can facilitate or hinder feedback-guided action regulation. We will discuss the main aspects of this potential mechanism in this section.

The Influence of Sense of Agency on Action Regulation

As illustrated in **Figure 1**, we assume a bidirectional relationship between subjective sense of agency and the objective ability to regulate one's actions and the environment. Under normal circumstances, being successful in regulating one's behavior according to one's current goals increases sense of agency (Moscarello and Hartley, 2017). Conversely, there is also evidence that subjective sense of agency can influence objective action regulation performance. Learned helplessness describes the phenomenon that the experience of having no control can lead to diminished performance in learning tasks (Maier and Seligman, 2016). Thus, low sense of control can have a detrimental effect on action regulation capacities.

Enhanced sense of agency has been related to better performance in tasks which require action regulation. For example, during motor learning tasks participants usually have to perform training sessions to learn challenging motor actions which require efficient or precise motor movements. Sense of agency during training can be induced by, for example, letting people choose the order of training tasks they have to perform. High vs. low sense of control has been found to lead to increased training success, meaning stronger improvements in motor performance (Sanli et al., 2013; Lewthwaite et al., 2015; Halperin et al., 2017; Iwatsuki et al., 2019; Iwatsuki and Otten, 2020; Matsumiya, 2021). Additionally, increased agency has been found to lead to lower error rates during cognitive control tasks (Legault and Inzlicht, 2013), and improved learning rates during memory tasks (Murayama et al., 2015; Murty et al., 2015). These findings suggest that sense of agency can facilitate action regulation.

The Influence of Affective Processing on Action Regulation

Action regulation is often related to the processing of affective information: we tend to alter our behavior when it leads to negative results and repeat the same actions when they are followed by positive outcomes. Thus, the monitoring of positive or negative performance feedback is vital for behavioral adjustments (Ullsperger et al., 2014). Negative feedback can lead to cognitive and neural changes, such as increased activity in brain circuits involved in cognitive control, which facilitate changes of ongoing behavior (van Driel et al., 2012; Beatty et al., 2020; Kaiser et al., 2021). For example, the affective-signaling theory proposes that affect is an important component of the neural conflict monitoring system, with negative affect eliciting an increase in executive control resources (Dignath et al., 2020). On the other hand, positive feedback is known to elicit increased activity in reward-related brain areas, which can lead to a reinforcement of goal-compatible behavior (Holroyd and Coles, 2002; Krigolson, 2018). Therefore, sensitivity to affective feedback is assumed to be one determinant of action regulation success (van de Vijver et al., 2011; Kaiser et al., 2021). Diminished sensitivity toward affective feedback has been associated with maladaptive behavior, such as diminished self-control in daily life (Overmeyer et al., 2021). Additionally, blunted neural reactivity toward errors or reward feedback has been observed in several psychological disorders which are characterized by self-regulatory problems, such as substance abuse or pathological gambling (Euser et al., 2013; Gorka et al., 2019; Li et al., 2020). Accordingly, any psychological factor that significantly modulates the impact of positive or negative feedback could potentially influence action regulation performance.

Affective Processing as a Mediator Between Sense of Agency and Action Regulation

As discussed above, several studies have found that high compared to low sense of agency is associated with changes in the processing of affective information, such as changes in the neural impact of affective feedback. Affective feedback is an important determinant of action regulation. Based on these findings, the model outlined in **Figure 1** assumes that affective processing mediates the influence of sense of agency on action regulation. The proposed relationship in **Figure 1** should be understood as a partial mediation, meaning that there are most likely other, non-affective mediators between agency experience and action regulation. For example, cognitive beliefs about one's self-efficacy might also partly predetermine regulative success (Sanli et al., 2013). Overall, we assume that increases in sense of agency lead to heightened neural sensitivity for affective feedback. Heightened sensitivity for affective feedback improves feedback learning and thus increases the chances to successfully self-regulate behavior. Conversely, low sense of agency could blunt sensitivity for affective feedback. This agency-induced decrease in feedback sensitivity could diminish feedback learning performance, thus having a detrimental influence on action regulation. Accordingly, affective processing could represent a specific mechanism which links subjective experience of agency with the objective ability to regulate one's behavior.

As discussed above, previous research provides ample evidence for a link between sense of agency and affective processing on the one hand (e.g., Leotti et al., 2010; Chambon et al., 2020; Zheng et al., 2020), and affective processing and action regulation on the other hand (e.g., Holroyd and Coles, 2002; Dignath et al., 2020; Kaiser et al., 2021). However, it should be noted that so far there are almost no empirical tests of the potential mediating role of affective feedback processing between sense of agency and action regulation. To the best of our knowledge only one study so far tested a closely related hypothesis: Legault and Inzlicht (2013) investigated the effect of feeling of autonomy on the performance in a cognitive control task. Autonomy was induced by providing an illusion of choice over the task, meaning that their operationalization of autonomy effectively manipulated choice agency. It was found that illusion of choice led to lower error rates, indicating improved action control. Importantly, the increase in performance for participants with choice agency was statistically mediated by stronger neural reactions toward error feedback, as measured via the feedback-related negativity component with EEG. It was concluded that increased error sensitivity mediates the relation between the feeling of autonomy and action control. This finding is consistent with the proposed interrelation between sense of agency, affective processing and action regulation.

Interestingly, Legault and Inzlicht (2013) found a mediation effect selectively for neural reactivity toward negative, but not positive feedback. This suggests that sense of agency influences action regulation by selective increases in error sensitivity. However, as discussed above, there are inconsistent results regarding sense of agency selectively boosting the processing of positive feedback (Mühlberger et al., 2017; Chambon et al., 2020), negative feedback (Bellebaum et al., 2010; Legault and Inzlicht, 2013), or both (Zheng et al., 2020). Accordingly, it remains an open question if the mediation of agency effects on regulation performance is primarily driven by changes in positive or negative feedback processing. Overall, due to the lack of more empirical reports regarding this question, the proposed mediating role of affective processing between sense of agency and action regulation remains tentative. However, we believe that investigating this link will be a promising avenue to develop a mechanistic understanding of the interaction between agency experience and goal-directed behavior.

CONCLUSIONS AND FUTURE DIRECTIONS

To summarize, experimental research indicates a bidirectional relation between sense of agency and affective processes. Several studies found evidence that emotional stimuli and/or affective states can, to a certain extent, have an influence on sense of agency. Conversely, manipulations of sense of agency have been shown to be associated with changes in affective states, as well as changes in the processing of affective information. Since the processing of affective information, particularly positive and negative performance feedback, is crucial for learning and action regulation, affective processing represents a potential link between the subjective feeling of being in control and the actual ability to gain control over one's actions and the environment. Our review has identified several questions which need to be clarified to fully understand and specify the bidirectional interrelation between sense of agency and affective processing, as well as its functional implications for action regulation.

For determining the influence of affective information on sense of agency, it would be important to clarify the discrepancy between affect-related effects on implicit measures of agency. While most studies employing self-report measures find that positive compared to negative affect increases sense of agency, experiment using implicit measures such as sensory attenuation or intentional binding come to diverging conclusions about the influence of affect on sense of agency. This suggests that emotional effects on implicit measures depend on additional variables which have not yet been clearly identified (but see Beck et al., 2017; Yoshie and Haggard, 2017). Importantly, it needs to be determined in how far affect-related effects on measures such as temporal binding and sensory attenuation indicate genuine alterations in sense of agency, rather than the susceptibility of implicit measures to perceptual or cognitive influences which do not directly reflect agency experience (Buehner, 2012; Kaiser and Schütz-Bosbach, 2018).

Concerning the influence of sense of agency on affective processing, future studies need to distinguish valence-independent from valence-specific effects. Numerous studies show that high compared to low sense of agency increase the neural impact of affective feedback. It is less clear in how far agency-induced changes in affective processing reflect a general increase in sensitivity for performance feedback as compared to a processing bias for either positive or negative feedback. If sense of agency selectively increased neural sensitivity for positive feedback, this could help to explain the neural underpinnings of the self-serving attributional bias, meaning increased sensitivity for positive results of self-determined actions (Mezulis et al., 2004; Chambon et al., 2020). Conversely, if sense of agency increased sensitivity for negative feedback, this could potentially represent an adaptive mechanism to adjust one's behavior after self-produced errors.

Lastly, more research is needed on the functional implications of the link between sense of agency and affective processing. Sense of agency can sometimes boost or diminish performance during goal-directed behavior. Since agency experience modulates the impact of affective feedback, and affective feedback is crucial for behavioral adjustments, affective processing is a promising candidate for a mediating factor between sense of agency and action regulation (Legault and Inzlicht, 2013). However, this possibility needs to be investigated empirically.

It will be important to clarify the neural mechanisms that link affective processing and sense of agency. Potential candidate mechanisms include limbic structures in the basal ganglia, such as the ventral striatum which is involved in the processing of reward, and areas of the medial prefrontal cortex, which are assumed to play a role in the processing self-relevant information (Cockburn et al., 2014; Wang and Delgado, 2019). Some studies indicate that intercommunication between these areas might be related to changes in sense of agency due to affective performance feedback (Wang and Delgado, 2019; Stolz et al., 2020). Moreover, it is noteworthy that both emotional processing and sense of agency have been related to the processing of bodily information. Affective stimulation is often accompanied by peripheral physiological changes (Kreibig, 2010). At the same time, sense of agency is assumed to be related to the sense of body ownership, meaning the feeling of having and controlling one's own body (Asai, 2015; Braun et al., 2018; Gonzalez-Franco et al., 2020). Since both agency experience and affective experience might partially rely on bodily information, the role of bodily changes in linking these two processes could be an important point of consideration in future studies.

The potential relationship between affective processing and sense of agency could have implications for psychopathological conditions that are marked by distortions in agency experience, such as schizophrenia or depersonalization disorder (van Haren et al., 2019; Kozáková et al., 2020). For example, schizophrenia has sometimes been linked to a distorted processing of affective information (Rahm et al., 2015; Maher et al., 2016). For such disorders, it would be important to know if alterations in agency experience and affective processing might be related. Lastly, previous research has separately investigated the developmental trajectory of affective processing (Quinn et al., 2011; Hoemann et al., 2019) and the evolving sense of agency (Zaadnoordijk et al., 2020; Meyer and Hunnius, 2021). With respect to their potential interactions, future research could probe the question in how far developmental changes in the sense of agency and affective processing co-occur or are even functionally related.

To conclude, while most previous research focusses on non-affective sensory and cognitive determinants of sense of agency, there are numerous findings which indicate that affective processes play an important role in our agency experience. Future research needs to specify the interactions between affect and sense of agency, for example with regards to valence-specificity of agency-related effects, as well as the role of other contextual factors which determine the influence of emotional information on agency experience. Importantly, studying the relation between sense of agency and affective processing could be a crucial step in linking the subjective experience of agency to failure or success during goal-directed behavior.

AUTHOR CONTRIBUTIONS

JK, MB, and SG reviewed the literature. JK drafted the manuscript. AG and SS-B contributed to the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a grant to AG and SS-B by the German Research Foundation (DFG)—Project No. 402781060.

REFERENCES

Aarts, H., Bijleveld, E., Custers, R., Dogge, M., Deelder, M., Schutter, D., et al. (2012). Positive priming and intentional binding: eye-blink rate predicts reward information effects on the sense of agency. *Soc. Neurosci.* 7, 105–112. doi: 10.1080/17470919.2011.590602

Abelson, J. L., Khan, S., Liberzon, I., Erickson, T. M., and Young, E. A. (2008). Effects of perceived control and cognitive coping on endocrine stress responses to pharmacological activation. *Biol. Psychiatry* 64, 701–707. doi: 10.1016/j.biopsych.2008.05.007

Asai, T. (2015). Illusory body-ownership entails automatic compensative movement: for the unified representation between body and action. *Exper. Brain Res.* 233, 777–85. doi: 10.1007/s00221-014-4153-0

Baess, P., Horváth, J., Jacobsen, T., and Schröger, E. (2011). Selective suppression of self-initiated sounds in an auditory stream: an ERP study. *Psychophysiology* 48, 1276–1283. doi: 10.1111/j.1469-8986.2011.01196.x

Barlas, Z., Hockley, W. E., and Obhi, S. S. (2017). The effects of freedom of choice in action selection on perceived mental effort and the sense of agency. *Acta Psychol.* 180, 122–129. doi: 10.1016/j.actpsy.2017.09.004

Barlas, Z., Hockley, W. E., and Obhi, S. S. (2018). Effects of free choice and outcome valence on the sense of agency: evidence from measures of intentional binding and feelings of control. *Exper. Brain Res.* 236, 129–139. doi: 10.1007/s00221-017-5112-3

Barlas, Z., and Obhi, S. S. (2014). Cultural background influences implicit but not explicit sense of agency for the production of musical tones. *Conscious. Cogn.* 28, 94–103. doi: 10.1016/j.concog.2014.06.013

Beatty, P. J., Buzzell, G. A., Roberts, D. M., and McDonald, C. G. (2020). Contrasting time and frequency domains: ERN and induced theta oscillations differentially predict post-error behavior. *Cognit. Affect. Behav. Neurosci.* 20, 636–647. doi: 10.3758/s13415-020-00792-7

Beck, B., Di Costa, S., and Haggard, P. (2017). Having control over the external world increases the implicit sense of agency. *Cognition* 162, 54–60. doi: 10.1016/j.cognition.2017.02.002

Bellebaum, C., and Colosio, M. (2014). From feedback- to response-based performance monitoring in active and observational learning. *J. Cogn. Neurosci.* 26, 2111–2127. doi: 10.1162/jocn_a_00612

Bellebaum, C., Kobza, S., Thiele, S., and Daum, I. (2010). It was not my fault: event-related brain potentials in active and observational learning from feedback. *Cereb. Cortex* 20, 2874–2883. doi: 10.1093/cercor/bhq038

Bhanji, J. P., and Delgado, M. R. (2014). Perceived control influences neural responses to setbacks and promotes persistence. *Neuron* 83, 1369–1375. doi: 10.1016/j.neuron.2014.08.012

Bhanji, J. P., Kim, E. S., and Delgado, M. R. (2016). Perceived control alters the effect of acute stress on persistence. *J. Exper. Psychol. Gen.* 145, 356–365. doi: 10.1037/xge0000137

Bismark, A. W., Hajcak, G., Whitworth, N. M., and Allen, J. J. B. (2013). The role of outcome expectations in the generation of the feedback-related negativity. *Psychophysiology* 50, 125–133. doi: 10.1111/j.1469-8986.2012.01490.x

Bjork, J. M., and Hommer, D. W. (2007). Anticipating instrumentally obtained and passively-received rewards: a factorial fMRI investigation. *Behav. Brain Res.* 177, 165–170. doi: 10.1016/j.bbr.2006.10.034

Blakemore, S. J., Wolpert, D. M., and Frith, C. (2000). Why can't you tickle yourself? *Neuroreport* 11, R11–R16. doi: 10.1586/14737175.7.10.1337

Blakemore, S. J., Wolpert, D. M., and Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nat. Neurosci.* 1, 635–640. doi: 10.1038/2870

Bobadilla-Suarez, S., Sunstein, C. R., and Sharot, T. (2017). The intrinsic value of choice: the propensity to under-delegate in the face of potential gains and losses. *J. Risk Uncertain.* 54, 187–202. doi: 10.1007/s11166-017-9259-x

Borhani, K., Beck, B., and Haggard, P. (2017). Choosing, doing, and controlling: implicit sense of agency over somatosensory events. *Psychol. Sci.* 28, 882–893. doi: 10.1177/0956797617697693

Bräscher, A. K., Becker, S., Hoeppli, M. E., and Schweinhardt, P. (2016). Different brain circuitries mediating controllable and uncontrollable pain. *J. Neurosci.* 36, 5013–5025. doi: 10.1523/JNEUROSCI.1954-15.2016

Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H. O., et al. (2018). The senses of agency and ownership: a review. *Front. Psychol.* 9:535. doi: 10.3389/fpsyg.2018.00535

Buehner, M. J. (2012). Understanding the past, predicting the future: causation, not intentional action, is the root of temporal binding. *Psychol. Sci.* 23, 1490–1497. doi: 10.1177/0956797612444612

Buehner, M. J., and May, J. (2003). Rethinking temporal contiguity and the judgement of causality: effects of prior knowledge, experience, and reinforcement procedure. *Q. J. Exp. Psychol. A* 56, 865–890. doi: 10.1080/02724980244000675

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., and Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nat. Hum. Behav.* 4, 1067–1079. doi: 10.1038/s41562-020-0919-5

Chernev, A., Böckenholt, U., and Goodman, J. (2012). Choice overload: a conceptual review and meta-analysis. *J. Consum. Psychol.* 25, 333–358. doi: 10.1016/j.jcps.2014.08.002

Christensen, J. F., Di Costa, S., Beck, B., and Haggard, P. (2019). I just lost it! Fear and anger reduce the sense of agency: a study using intentional binding. *Exper. Brain Res.* 237, 1205–1212. doi: 10.1007/s00221-018-5461-6

Christensen, J. F., Yoshie, M., Di Costa, S., and Haggard, P. (2016). Emotional valence, sense of agency and responsibility: a study using intentional binding. *Conscious. Cogn.* 43, 1–10. doi: 10.1016/j.concog.2016.02.016

Cockburn, J., Collins, A. G. E., and Frank, M. J. (2014). A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron* 83, 551–557. doi: 10.1016/j.neuron.2014.06.035

Craig, A.R., Franklin, J. A., and Andrews, G. (1984). A scale to measure locus of control of behaviour. *Br. J. Med. Psychol.* 57, 173–180. doi: 10.1111/j.2044-8341.1984.tb01597.x

Di Costa, S., Théro, H., Chambon, V., and Haggard, P. (2018). Try and try again: post-error boost of an implicit measure of agency. *Q. J. Exper. Psychol.* 71, 1584–1595. doi: 10.1080/17470218.2017.1350871

Dignath, D., Eder, A. B., Steinhauser, M., and Kiesel, A. (2020). Conflict monitoring and the affective-signaling hypothesis—an integrative review. *Psychonomic Bull. Rev.* 27, 193–216. doi: 10.3758/s13423-019-01668-9

Eitam, B., Kennedy, P. M., and Higgins, E. T. (2013). Motivation from control. *Exper. Brain Res.* 229, 475–484. doi: 10.1007/s00221-012-3370-7

Euser, A. S., Evans, B. E., Greaves-Lord, K., Huizink, A. C., and Franken, I. H. A. (2013). Diminished error-related brain activity as a promising endophenotype for substance-use disorders: evidence from high-risk offspring. *Addict. Biol.* 18, 970–984. doi: 10.1111/adb.12002

Fang, H., Wan, X., Zheng, S., and Meng, L. (2020). The spillover effect of autonomy frustration on human motivation and its electrophysiological representation. *Front. Hum. Neurosci.* 14:134. doi: 10.3389/fnhum.2020.00134

Fujiwara, J., Usui, N., Park, S. Q., Williams, T., Iijima, T., Taira, M., et al. (2013). Value of freedom to choose encoded by the human brain. *J. Neurophysiol.* 110, 1915–1929. doi: 10.1152/jn.01057.2012

Gallagher, M. W., Bentley, K. H., and Barlow, D. H. (2014). Perceived control and vulnerability to anxiety disorders: a meta-analytic review. *Cognit. Ther. Res.* 38, 571–584. doi: 10.1007/s10608-014-9624-x

Galvin, B. M., Randel, A. E., Collins, B. J., and Johnson, R. E. (2018). Changing the focus of locus (of control): a targeted review of the locus of control literature and agenda for future research. *J. Organ. Behav.* 39, 820–833. doi: 10.1002/job.2275

Gentsch, A., and Schütz-Bosbach, S. (2015). "Agency and outcome prediction," in *The Sense of Agency*, eds P. Haggard, and B. Eitam (Oxoford: Oxford University Press), 217–234. doi: 10.1093/acprof:oso/9780190267278.003.0009

Gentsch, A., Schütz-Bosbach, S., Endrass, T., and Kathmann, N. (2012). Dysfunctional forward model mechanisms and aberrant sense of agency in obsessive-compulsive disorder. *Biol. Psychiatry* 71, 652–659. doi: 10.1016/j.biopsych.2011.12.022

Gentsch, A., and Synofzik, M. (2014). Affective coding: the emotional dimension of agency. *Front. Hum. Neurosci.* 8:608. doi: 10.3389/fnhum.2014.00608

Gentsch, A., Weiss, C., Spengler, S., Synofzik, M., and Schütz-Bosbach, S. (2015). Doing good or bad: how interactions between action and emotion expectations shape the sense of agency. *Soc. Neurosci.* 10, 418–30. doi: 10.1080/17470919.2015.1006374

Gonzalez-Franco, M., Cohn, B., Ofek, E., Burin, D., and Maselli, A. (2020). "The self-avatar follower effect in virtual reality," in *Proceedings - 2020 IEEE Conference on Virtual Reality and 3D User Interfaces* (Atlanta, GA). doi: 10.1109/VR46266.2020.1580500165557

Gorka, S. M., Lieberman, L., Kreutzer, K. A., Carrillo, V., Weinberg, A., and Shankman, S. A. (2019). Error-related neural activity and alcohol use disorder: differences from risk to remission. *Progr. Neuro Psychopharmacol. Biol. Psychiatry* 92, 271–278. doi: 10.1016/j.pnpbp.2019.01.011

Gratton, G., Cooper, P., Fabiani, M., Carter, C. S., and Karayanidis, F. (2017). Dynamics of cognitive control: theoretical bases, paradigms, and a view for the future. *Psychophysiology* 55, 1–29. doi: 10.1111/psyp.13016

Greifeneder, R., Scheibehenne, B., and Kleber, N. (2010). Less may be more when choosing is difficult: choice complexity and too much choice. *Acta Psychol.* 133, 45–50. doi: 10.1016/j.actpsy.2009.08.005

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 197–208. doi: 10.1038/nrn.2017.14

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nat. Neurosci.* 5, 382–385. doi: 10.1038/nn827

Haggard, P., and Tsakiris, M. (2009). The experience of agency. *Curr. Dir. Psychol. Sci.* 18, 242–246. doi: 10.1111/j.1467-8721.2009.01644.x

Halperin, I., Chapman, D. W., Martin, D. T., Lewthwaite, R., and Wulf, G. (2017). Choices enhance punching performance of competitive kickboxers. *Psychol. Res.* 81, 1051–1058. doi: 10.1007/s00426-016-0790-1

Harnett, N. G., Wheelock, M. D., Wood, K. H., Ladnier, J. C., Mrug, S., and Knight, D. C. (2015). Affective state and locus of control modulate the neural response to threat. *Neuroimage* 121, 217–226. doi: 10.1016/j.neuroimage.2015.07.034

Hassall, C. D., Hajcak, G., and Krigolson, O. E. (2019). The importance of agency in human reward processing. *Cognit. Affect. Behav. Neurosci.* 19, 1458–1466. doi: 10.3758/s13415-019-00730-2

Herman, A. M., and Tsakiris, M. (2020). Feeling in control: the role of cardiac timing in the sense of agency. *Affect. Sci.* 1, 155–171. doi: 10.1007/s42761-020-00013-x

Hoemann, K., Xu, F., and Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: a constructionist hypothesis. *Dev. Psychol.* 55, 1830–1849. doi: 10.1037/dev0000686

Holroyd, C. B., and Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* 109, 679–709. doi: 10.1037/0033-295X.109.4.679

Iwatsuki, T., and Otten, M. P. (2020). Providing choice enhances motor performance under psychological pressure. *J. Motor Behav.* 53, 656–666. doi: 10.1080/00222895.2020.1833827

Iwatsuki, T., Shih, H. T., Abdollahipour, R., and Wulf, G. (2019). More bang for the buck: autonomy support increases muscular efficiency. *Psychol. Res.* 85, 439–445. doi: 10.1007/s00426-019-01243-w

Iyengar, S. S., and Lepper, M. R. (2000). When choice is demotivating: can one desire too much of a good thing? *J. Pers. Soc. Psychol.* 79, 995–1006. doi: 10.1037/0022-3514.79.6.995

Kaiser, J., Belenya, R., Chung, W.-Y., Gentsch, A., and Schütz-Bosbach, S. (2021). Learning something new versus changing your ways: distinct effects on midfrontal oscillations and cardiac activity for learning and flexible adjustments. *Neuroimage* 226:117550. doi: 10.1016/j.neuroimage.2020.117550

Kaiser, J., and Schütz-Bosbach, S. (2018). Sensory attenuation of self-produced signals does not rely on self-specific motor predictions. *Eur. J. Neurosci.* 47, 1303–1310. doi: 10.1111/ejn.13931

Kaiser, J., and Schütz-Bosbach, S. (2019). Proactive control without midfrontal control signals? The role of midfrontal oscillations in preparatory conflict adjustments. *Biol. Psychol.* 148:107747. doi: 10.1016/j.biopsycho.2019.107747

Kaiser, J., and Schütz-Bosbach, S. (2021). Motor interference, but not sensory interference, increases midfrontal theta activity and brain synchronization during reactive control. *J. Neurosci.* 41, 1788–1801. doi: 10.1523/JNEUROSCI.1682-20.2020

Kaiser, J., Simon, N. A., Sauseng, P., and Schütz-Bosbach, S. (2019). Midfrontal neural dynamics distinguish between general control and inhibition-specific processes in the stopping of motor actions. *Sci. Rep.* 9:13054. doi: 10.1038/s41598-019-49476-4

Karsh, N., and Eitam, B. (2015). I control therefore I do: judgments of agency influence action selection. *Cognition* 138, 122–131. doi: 10.1016/j.cognition.2015.02.002

Koffer, R., Drewelies, J., Almeida, D. M., Conroy, D. E., Pincus, A. L., Gerstorf, D., et al. (2019). The role of general and daily control beliefs for affective stressor-reactivity across adulthood and old age. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 74, 242–253. doi: 10.1093/geronb/gbx055

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., and De Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. *Cereb. Cortex* 22, 2197–2206. doi: 10.1093/cercor/bhr310

Kozáková, E., Bakštein, E., Havlíček, O., Bečev, O., Knytl, P., Zaytseva, Y., et al. (2020). Disrupted sense of agency as a state marker of first-episode schizophrenia: a large-scale follow-up study. *Front. Psychiatry* 11:570570. doi: 10.3389/fpsyt.2020.570570

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biol. Psychol.* 84, 394–421. doi: 10.1016/j.biopsycho.2010.03.010

Krigolson, O. E. (2018). Event-related brain potentials and the study of reward processing: methodological considerations. *Int. J. Psychophysiol.* 132, 175–183. doi: 10.1016/j.ijpsycho.2017.11.007

Kulakova, E., Khalighinejad, N., and Haggard, P. (2017). I could have done otherwise: availability of counterfactual comparisons informs the sense of agency. *Conscious. Cogn.* 49, 237–244. doi: 10.1016/j.concog.2017.01.013

Le Bars, S., Devaux, A., Nevidal, T., Chambon, V., and Pacherie, E. (2020). Agents' pivotality and reward fairness modulate sense of agency in cooperative joint action. *Cognition* 195:104117. doi: 10.1016/j.cognition.2019.104117

Legault, L., and Inzlicht, M. (2013). Self-determination, self-regulation, and the brain: autonomy improves performance by enhancing neuroaffective responsiveness to self-regulation failure. *J. Pers. Soc. Psychol.* 105, 123–138. doi: 10.1037/a0030426

Leotti, L. A., and Delgado, M. R. (2011). The inherent reward of choice. *Psychol. Sci.* 22, 1310–1318. doi: 10.1177/0956797611417005

Leotti, L. A., and Delgado, M. R. (2014). The value of exercising control over monetary gains and losses. *Psychol. Sci.* 25, 596–604. doi: 10.1177/0956797613514589

Leotti, L. A., Iyengar, S. S., and Ochsner, K. N. (2010). Born to choose: the origins and value of the need for control. *Trends Cogn. Sci.* 14, 457–463. doi: 10.1016/j.tics.2010.08.001

Lewthwaite, R., Chiviacowsky, S., Drews, R., and Wulf, G. (2015). Choose to move: the motivational impact of autonomy support on motor learning. *Psychonomic Bull. Rev.* 22, 1383–1388. doi: 10.3758/s13423-015-0814-7

Li, P., Han, C., Lei, Y., Holroyd, C. B., and Li, H. (2011). Responsibility modulates neural mechanisms of outcome processing: an ERP study. *Psychophysiology* 48, 1129–1133. doi: 10.1111/j.1469-8986.2011.01182.x

Li, Q., Wang, Y., Yang, Z., Dai, W., Zheng, Y., Sun, Y., et al. (2020). Dysfunctional cognitive control and reward processing in adolescents with Internet gaming disorder. *Psychophysiology* 57:e13469. doi: 10.1111/psyp.13469

Li, T., Zhao, F., and Yu, G. (2021). Who is more utilitarian? Negative affect mediates the relation between control deprivation and moral judgment. *Curr. Psychol.* 40, 4024–4030. doi: 10.1007/s12144-019-00301-1

Lorenz, R. C., Gleich, T., Kühn, S., Pöhland, L., Pelz, P., Wüstenberg, T., et al. (2015). Subjective illusion of control modulates striatal reward anticipation in adolescence. *Neuroimage* 117, 250–257. doi: 10.1016/j.neuroimage.2015.05.024

Ly, V., Wang, K. S., Bhanji, J., and Delgado, M. R. (2019). A reward-based framework of perceived control. *Front. Neurosci.* 13:65. doi: 10.3389/fnins.2019.00065

Maher, S., Ekstrom, T., and Chen, Y. (2016). Impaired visual cortical processing of affective facial information in schizophrenia. *Clin. Psychol. Sci.* 4, 651–660. doi: 10.1177/2167702615609595

Maier, S. F., and Seligman, M. E. P. (2016). Learned helplessness at fifty: insights from neuroscience. *Psychol. Rev.* 123, 349–367. doi: 10.1037/rev0000033

Majchrowicz, B., Kulakova, E., Di Costa, S., and Haggard, P. (2020). Learning from informative losses boosts the sense of agency. *Q. J. Exper. Psychol.* 73, 2272–2289. doi: 10.1177/1747021820958258

Majchrowicz, B., and Wierzchoń, M. (2021). Sensory attenuation of action outcomes of varying amplitude and valence. *Conscious. Cogn.* 87:103058. doi: 10.1016/j.concog.2020.103058

Martin, L. E., and Potts, G. F. (2011). Medial frontal event-related potentials and reward prediction: do responses matter? *Brain Cogn.* 77, 128–134. doi: 10.1016/j.bandc.2011.04.001

Matsumiya, K. (2021). Awareness of voluntary action, rather than body ownership, improves motor control. *Sci. Rep.* 11:418. doi: 10.1038/s41598-020-79910-x

Mei, S., Yi, W., Zhou, S., Liu, X., and Zheng, Y. (2018). Contextual valence modulates the effect of choice on incentive processing. *Soc. Cogn. Affect. Neurosci.* 13, 1249–1258. doi: 10.1093/scan/nsy098

Meng, L., and Ma, Q. (2015). Live as we choose: the role of autonomy support in facilitating intrinsic motivation. *Int. J. Psychophysiol.* 98, 441–447. doi: 10.1016/j.ijpsycho.2015.08.009

Meyer, M., and Hunnius, S. (2021). Neural processing of self-produced and externally generated events in 3-month-old infants. *J. Exp. Child Psychol.* 204:105039. doi: 10.1016/j.jecp.2020.105039

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., and Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychol. Bull.* 130, 711–747. doi: 10.1037/0033-2909.130.5.711

Mistry, P., and Liljeholm, M. (2016). Instrumental divergence and the value of control. *Sci. Rep.* 6:36295. doi: 10.1038/srep36295

Mohr, C., Leyendecker, S., and Helmchen, C. (2012). Effects of perceived and exerted pain control on neural activity during pain relief in experimental heat hyperalgesia: a fMRI study. *Eur. J. Pain* 16, 496–508. doi: 10.1016/j.ejpain.2011.07.010

Moore, J. W. (2016). What is the sense of agency and why does it matter. *Front. Psychol.* 7:1272doi: 10.3389/fpsyg.2016.01272

Moreton, J., Callan, M. J., and Hughes, G. (2017). How much does emotional valence of action outcomes affect temporal binding? *Conscious. Cogn.* 49, 25–34. doi: 10.1016/j.concog.2016.12.008

Moscarello, J. M., and Hartley, C. A. (2017). Agency and the calibration of motivated behavior. *Trends Cogn. Sci.* 21, 725–735. doi: 10.1016/j.tics.2017.06.008

Mühlberger, C., Angus, D. J., Jonas, E., Harmon-Jones, C., and Harmon-Jones, E. (2017). Perceived control increases the reward positivity and stimulus preceding negativity. *Psychophysiology* 54, 310–322. doi: 10.1111/psyp.12786

Murayama, K., Matsumoto, M., Izuma, K., Sugiura, A., Ryan, R. M., Deci, E. L., et al. (2015). How self-determined choice facilitates performance: a key role of the ventromedial prefrontal cortex. *Cereb. Cortex* 25, 1241–1251. doi: 10.1093/cercor/bht317

Murty, V. P., DuBrow, S., and Davachi, L. (2015). The simple act of choosing influences declarative memory. *J. Neurosci.* 35, 6255–6264. doi: 10.1523/JNEUROSCI.4181-14.2015

Nataraj, R., Hollinger, D., Liu, M., and Shah, A. (2020). Disproportionate positive feedback facilitates sense of agency and performance for a reaching movement task with a virtual hand. *PLoS ONE* 15:e0233175. doi: 10.1371/journal.pone.0233175

Obhi, S. S., Swiderski, K. M., and Farquhar, R. (2013). Activating memories of depression alters the experience of voluntary action. *Exp. Brain Res.* 229, 497–506. doi: 10.1007/s00221-012-3372-5

Oishi, H., Tanaka, K., and Watanabe, K. (2018). Feedback of action outcome retrospectively influences sense of agency in a continuous action task. *PLoS ONE* 13:e0202690. doi: 10.1371/journal.pone.0202690

Oishi, H., Tanaka, K., and Watanabe, K. (2019). Sense of agency in continuous action is influenced by outcome feedback in one-back trials. *Acta Psychol.* 199:102897. doi: 10.1016/j.actpsy.2019.102897

Orgaz, C., Estévez, A., and Matute, H. (2013). Pathological gamblers are more vulnerable to the illusion of control in a standard associative learning task. *Front. Psychol.* 4:306. doi: 10.3389/fpsyg.2013.00306

Osumi, T., Tsuji, K., Shibata, M., and Umeda, S. (2019). Machiavellianism and early neural responses to others' facial expressions caused by one's own decisions. *Psychiatry Res.* 271, 669–677. doi: 10.1016/j.psychres.2018.12.037

Overmeyer, R., Berghäuser, J., Dieterich, R., Wolff, M., Goschke, T., and Endrass, T. (2021). The error-related negativity predicts self-control failures in daily life. *Front. Hum. Neurosci.* 14:614979. doi: 10.3389/fnhum.2020.614979

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

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

Proudfit, G. H. (2015). The reward positivity: from basic research on reward to a biomarker for depression. *Psychophysiology* 52, 449–59. doi: 10.1111/psyp.12370

Quinn, P. C., Anzures, G., Izard, C. E., Lee, K., Pascalis, O., Slater, A. M., et al. (2011). Looking across domains to understand infant representation of emotion. *Emot. Rev.* 3, 197–206. doi: 10.1177/1754073910387941

Rahm, C., Liberg, B., Reckless, G., Ousdal, O., Melle, I., Andreassen, O. A., et al. (2015). Negative symptoms in schizophrenia show association with amygdala volumes and neural activation during affective processing. *Acta Neuropsychiatr.* 27, 213–220. doi: 10.1017/neu.2015.11

Reutskaja, E., and Hogarth, R. M. (2009). Satisfaction in choice as a function of the number of alternatives: when "goods satiate." *Psychol. Mark.* 26, 197–203. doi: 10.1002/mar.20268

Romaniuk, L., Sandu, A. L., Waiter, G. D., McNeil, C. J., Xueyi, S., Harris, M. A., et al. (2019). The neurobiology of personal control during reward learning and its relationship to mood. *Biol. Psychiatry Cognit. Neurosci. Neuroimaging* 4, 190–199. doi: 10.1016/j.bpsc.2018.09.015

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145–172. doi: 10.1037/0033-295X.110.1.145

Salomons, T., Detloff, A. M., and Johnstone, T. (2014). Neural emotion regulation circuitry underlying anxiolytic effects of perceived control over pain. *J. Cogn. Neurosci.* 27, 222–233. doi: 10.1162/jocn_a_00702

Salomons, T. V., Johnstone, T., Backonja, M. M., and Davidson, R. J. (2004). Perceived controllability modulates the neural response to pain. *J. Neurosci.* 24, 7199–7203. doi: 10.1523/JNEUROSCI.1315-04.2004

Sanli, E. A., Patterson, J. T., Bray, S. R., and Lee, T. D. (2013). Understanding self-controlled motor learning protocols through the self-determination theory. *Front. Psychol.* 3:611. doi: 10.3389/fpsyg.2012.00611

Scheibehenne, B., Greifeneder, R., and Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *J. Consum. Res.* 37, 409–425. doi: 10.1086/651235

Stern, Y., Koren, D., Moebus, R., Panishev, G., and Salomon, R. (2020). Assessing the relationship between sense of agency, the bodily-self and stress: four virtual-reality experiments in healthy individuals. *J. Clin. Med.* 9:2931. doi: 10.3390/jcm9092931

Stolz, D. S., Müller-Pinzler, L., Krach, S., and Paulus, F. M. (2020). Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nat. Commun.* 11:1230. doi: 10.1038/s41467-020-14800-4

Studer, B., Geniole, S. N., Becker, M. L., Eisenegger, C., and Knecht, S. (2020). Inducing illusory control ensures persistence when rewards fade and when others outperform us. *Psychonomic Bull. Rev.* 27, 809–818. doi: 10.3758/s13423-020-01745-4

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* 17, 219–239. doi: 10.1016/j.concog.2007.03.010

Synofzik, M., Vosgerau, G., and Voss, M. (2013). The experience of agency: an interplay between prediction and postdiction. *Front. Psychol.* 4:127. doi: 10.3389/fpsyg.2013.00127

Szczepanowski, R., Traczyk, J., Wierzchoń, M., and Cleeremans, A. (2013). The perception of visual emotion: comparing different measures of awareness. *Conscious. Cogn.* 22, 212–220. doi: 10.1016/j.concog.2012.12.003

Takahata, K., Takahashi, H., Maeda, T., Umeda, S., Suhara, T., Mimura, M., et al. (2012). It's not my fault: postdictive modulation of intentional binding by monetary gains and losses. *PLoS ONE* 7:e53421. doi: 10.1371/journal.pone.0053421

Tanaka, T., Watanabe, K., and Tanaka, K. (2020). Immediate action effects motivate actions based on the stimulus–response relationship. *Exper. Brain Res.* 68, 138–154. doi: 10.1007/s00221-020-05955-z

Thuillard, S., and Dan-Glauser, E. S. (2017). The regulatory effect of choice in situation selection reduces experiential, exocrine and respiratory arousal for negative emotional stimulations. *Sci. Rep.* 7:12626. doi: 10.1038/s41598-017-12626-7

Thuillard, S., and Dan-Glauser, E. S. (2020). Efficiency of illusory choice used as a variant of situation selection for regulating emotions: reduction of positive experience but preservation of physiological downregulation. *Appl. Psychophysiol. Biofeedback* 46, 115–132. doi: 10.1007/s10484-020-09484-x

Timm, J., Schönwiesner, M., Schröger, E., and SanMiguel, I. (2016). Sensory suppression of brain responses to self-generated sounds is observed with and without the perception of agency. *Cortex* 80, 5–20. doi: 10.1016/j.cortex.2016.03.018

Tricomi, E. M., Delgado, M. R., and Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron* 41, 281–292. doi: 10.1016/S0896-6273(03)00848-1

Ullsperger, M., Danielmeier, C., and Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiol. Rev.* 94, 35–79. doi: 10.1152/physrev.00041.2012

van de Vijver, I., Ridderinkhof, K. R., and Cohen, M. X. (2011). Frontal oscillatory dynamics predict feedback learning and action adjustment. *J. Cogn. Neurosci.* 23, 4106–4121. doi: 10.1162/jocn_a_00110

van der Wel, R. P. R. D., Sebanz, N., and Knoblich, G. (2012). The sense of agency during skill learning in individuals and dyads. *Conscious. Cogn.* 21, 1267–1279. doi: 10.1016/j.concog.2012.04.001

van Driel, J., Ridderinkhof, K. R., and Cohen, M. X. (2012). Not all errors are alike: theta and alpha EEG dynamics relate to differences in error-processing dynamics. *J. Neurosci.* 32, 16795–16806. doi: 10.1523/JNEUROSCI.0802-12.2012

van Haren, N., van der Weiden, A., Aarts, H., and Prikken, M. (2019). Sense of ownership and sense of agnecy in schizophrenia patients. *Schizophr. Bull.* 45(Suppl. 2):S107. doi: 10.1093/schbul/sbz022.046

Vancleef, L. M. G., and Peters, M. L. (2011). The influence of perceived control and self-efficacy on the sensory evaluation of experimentally induced pain. *J. Behav. Ther. Exp. Psychiatry* 42, 511–517. doi: 10.1016/j.jbtep.2011.05.006

Wang, K. S., and Delgado, M. R. (2019). Corticostriatal circuits encode the subjective value of perceived control. *Cereb. Cortex* 29, 5049–5060. doi: 10.1093/cercor/bhz045

Weiss, C., Herwig, A., and Schütz-Bosbach, S. (2011). The self in action effects: selective attenuation of self-generated sounds. *Cognition* 121, 207–218. doi: 10.1016/j.cognition.2011.06.011

Wiech, K., Kalisch, R., Weiskopf, N., Pleger, B., Stephan, K. E., and Dolan, R. J. (2006). Anterolateral prefrontal cortex mediates the analgesic effect of expected and perceived control over pain. *J. Neurosci.* 26, 11501–11509. doi: 10.1523/JNEUROSCI.2568-06.2006

Wolpe, N., and Rowe, J. B. (2014). Beyond the "urge to move": objective measures for the study of agency in the post-Libet era. *Front. Hum. Neurosci.* 8:450. doi: 10.3389/fnhum.2014.00450

Yeung, N., Holroyd, C. B., and Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cereb. Cortex* 15, 535–544. doi: 10.1093/cercor/bhh153

Yi, W., Mei, S., Li, Q., Liu, X., and Zheng, Y. (2018). How choice influences risk processing: an ERP study. *Biol. Psychol.* 138, 223–230. doi: 10.1016/j.biopsycho.2018.08.011

Yoshie, M., and Haggard, P. (2013). Negative emotional outcomes attenuate sense of agency over voluntary actions. *Curr. Biol.* 23, 2028–2032. doi: 10.1016/j.cub.2013.08.034

Yoshie, M., and Haggard, P. (2017). Effects of emotional valence on sense of agency require a predictive model. *Sci. Rep.* 7:8733. doi: 10.1038/s41598-017-08803-3

Zaadnoordijk, L., Meyer, M., Zaharieva, M., Kemalasari, F., van Pelt, S., and Hunnius, S. (2020). From movement to action: an EEG study into the emerging sense of agency in early infancy. *Dev. Cogn. Neurosci.* 42:100760. doi: 10.1016/j.dcn.2020.100760

Zheng, Y., Wang, M., Zhou, S., and Xu, J. (2020). Functional heterogeneity of perceived control in feedback processing. *Soc. Cogn. Affect. Neurosci.* 15, 329–336. doi: 10.1093/scan/nsaa028

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

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

frontiers in Psychology

MINI REVIEW
published: 12 October 2021
doi: 10.3389/fpsyg.2021.716950

An Embodied Cognition Perspective on the Role of Interoception in the Development of the Minimal Self

Lisa Musculus[1], *Markus R. Tünte*[2], *Markus Raab*[1,3] *and Ezgi Kayhan*[4,5]*

[1]Institute of Psychology, German Sport University Cologne, Cologne, Germany, [2]Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria, [3]School of Applied Sciences, London South Bank University, London, United Kingdom, [4]Department of Developmental Psychology, University of Potsdam, Potsdam, Germany, [5]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Interoception is an often neglected but crucial aspect of the human minimal self. In this perspective, we extend the embodiment account of interoceptive inference to explain the development of the minimal self in humans. To do so, we first provide a comparative overview of the central accounts addressing the link between interoception and the minimal self. Grounding our arguments on the embodiment framework, we propose a bidirectional relationship between motor and interoceptive states, which jointly contribute to the development of the minimal self. We present empirical findings on interoception in development and discuss the role of interoception in the development of the minimal self. Moreover, we make theoretical predictions that can be tested in future experiments. Our goal is to provide a comprehensive view on the mechanisms underlying the minimal self by explaining the role of interoception in the development of the minimal self.

Keywords: interoception, bodily self, embodied cognition, cardioception, development of minimal self

OPEN ACCESS

Edited by:
Claudia Scorolli,
University of Bologna, Italy

Reviewed by:
Hari Narayanan V.,
Indian Institute of Technology,
Jodhpur, India
Thomas A. Stoffregen,
University of Minnesota, Twin Cities,
United States

Correspondence:
Ezgi Kayhan
kayhan@uni-potsdam.de

Specialty section:
This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

Received: *01 June 2021*
Accepted: *06 September 2021*
Published: *12 October 2021*

Citation:
Musculus L, Tünte MR, Raab M and Kayhan E (2021) An Embodied Cognition Perspective on the Role of Interoception in the Development of the Minimal Self.
Front. Psychol. 12:716950.
doi: 10.3389/fpsyg.2021.716950

INTEROCEPTION AND THE BODILY MINIMAL SELF

Body representation in humans is subsumed under the so-called bodily or minimal self, which is defined as a "person's phenomenal experience in the here and now" (Hafner et al., 2017, p. 1; Gallagher, 2000). The bodily or minimal self of humans is heavily dependent on the "embedded body" (Gallagher, 2000, p. 15). The minimal self consists of the sense of ownership, which refers to the feeling that one's body belongs to oneself, and the sense of agency, which is the feeling that one's actions cause effects (Gallagher, 2000; Verschoor and Hommel, 2017). Given the crucial role of the body in conceptualizing the sense of ownership and the sense of agency, and hence the human minimal self, it is surprising that *internal* bodily signals such as heartbeat and respiration have been largely ignored in this line of research (Tsakiris et al., 2011; Marshall et al., 2018; Seth and Tsakiris, 2018). For instance, a newborn's heart beats at ca. 127 beats per minute (bpm) increasing to a maximum of ca. 145 bpm within 1 month, before it decreases to 112 bpm by the age of 2 years (Fleming et al., 2011). Heartbeat perception is central to research on interoception, which is traditionally defined as the perception and sensation of the internal bodily signals (Murphy et al., 2017). From an embodied cognition perspective, it seems implausible that such bodily changes during development would not affect the body representation, hence the minimal self. In this paper, we argue that interoceptive signals are fundamental to the phenomenal experience of here and now constructing the minimal self. Grounding our arguments on the embodiment framework, we discuss how interoception shapes the development of the minimal self in humans.

Our perspective aims to extend the embodied cognition account of interoceptive inference (Marshall et al., 2018) by explicitly focusing on the role of interoception in the *development* of the human minimal self. The call for the research topic postulates embodied cognition as a powerful framework in explaining the minimal self (Hafner et al., 2021). Embodied cognition accounts are manifold (for detailed overviews see, e.g., Wilson, 2002; Shapiro, 2019) varying regarding central assumption and their "radicalism" (Raab and Araujo, 2019, p. 1) with respect to whether the link between the environment and perception, cognition and action is direct (e.g., Gibson, 1979; Chemero, 2011; Jacob, 2016) or mediated through representations (e.g., Newen et al., 2018). We base our perspective on the central assumption that representations benefit human's flexible and adaptive way of acting in a complex world (Schulz, 2018). We thereby take a "moderate" position (cf., Goldman, 2012), acknowledging at the same time that other approaches exist aiming to overcome the separation of approaches (e.g., Witt and Riley, 2014; Ciaunica et al., 2021). In addition, our embodied cognition perspective considers bodily changes relevant to explaining human development (Musculus et al., 2021), and, here, relate it to the development of the self.

Our contribution consists of a comparative overview of the central theoretical accounts explaining the link between interoception and the bodily minimal self (Marshall et al., 2018; Seth and Tsakiris, 2018). Based on this comparison, we present our embodied cognition perspective in more detail focusing on the emerging minimal self. Following a discussion of empirical findings on how interoception shapes the development of the bodily minimal self, we will outline theoretical predictions and a research program to better understand the role of interoception in the development of the bodily minimal self from our embodied cognition perspective.

Comparative Overview of Theoretical Accounts

In this part, we will compare two theoretical accounts that explain the role of interoception in the bodily minimal self on three levels (i.e., origin, central model assumptions, relation of interoception to the self).

The instrumental interoceptive inference account, proposed by Seth and Tsakiris (2018), originates from cybernetics and the free-energy principle. According to the instrumental interoceptive inference account (motor) actions serve the regulation of interoceptive states through a hierarchically organized generative model (Seth et al., 2012; Seth and Tsakiris, 2018): The generative model encodes priors of sensory information in higher levels of the neural hierarchy, based on which lower-level information such as interoceptive states are predicted. These top-down predictions are compared to the perceived interoceptive states. The difference between the predicted and the perceived states results in prediction errors, which are then sent back to the higher levels in the hierarchy to further update the generative models (Seth et al., 2012; Seth and Tsakiris, 2018). Through repetition of this hierarchical cascading, interoceptive prediction errors are minimized, which eventually maximizes the interoceptive generative models. These models form the basis of a sense of self and the experience of selfhood (Seth et al., 2012). Importantly, interoceptive prediction errors can also be minimized through action, also known as active inference. In the case of interoception, this refers to "intero-actions" (e.g., reflexes). Together, interoceptive (active) inference serves the overall goal of allostasis: maintaining physiological parameters of the body within a constant range by adapting to environmental change (Sterling, 2014; Seth and Tsakiris, 2018). This notion draws the connection to the experience of selfhood: Interoception fosters the stability of the bodily minimal self as opposed to the ever-changing exteroceptive information (Tsakiris, 2017).

Marshall et al. (2018) built up on this account and elaborated further on the functional link between interoception and (motor) actions. This approach is strongly influenced by cognitive psychology and cognitive neuroscience. According to Marshall et al. (2018), both motor and interoceptive states can form predictions about each other. Predictions are then compared to afferent, sensory input stemming from the sensorimotor system in the case of the motor prediction, and the autonomic system in the case of interoceptive predictions. Importantly, motor and interoceptive predictions are weighed equally in how they contribute to subjective experience emphasizing a functional bidirectional link. This also draws the connection to the experience of selfhood: Interoceptive states modulate the experience of selfhood just as strongly as (motor) actions (Marshall et al., 2018).

Embodiment Suggests a Bidirectional Link

Both theoretical accounts, although originating from different domains, share the idea that predictive coding can be considered as the "mechanistic process […] forming an initial, theoretical link between" (Marshall et al., 2018, p. 2) interoception and the minimal self. The accounts differ in how they elaborate on the functional relationship between interoception and motor processes. From an instrumental interoceptive inference account, the impact of motor predictions on interoceptive states has been formulated in terms of a hierarchically organized generative model (Seth et al., 2012). This was extended theoretically by explicitly suggesting a bidirectional link in which interoceptive states also predict motor actions (Marshall et al., 2018). We find the theoretical argument of bidirectionality plausible in line with the general tenets of the embodied cognition perspective.

Although both theoretical accounts mention and acknowledge the relevance of a developmental approach, neither of them focus on development in more detail. We tap into this gap and discuss the *development of the minimal self*. Recent reviews on this topic studied the development of the minimal self through experiencing and interacting with the external world (Georgie et al., 2019; Nguyen et al., 2021). We extend this line of research by considering the role of interoception in the development of the minimal self. In particular, we derive theoretical predictions on the developmental trajectory of interoception and discuss its relation to minimal-self dimensions such as the sense of ownership and agency from our embodied cognition perspective. To do so, we summarize the evidence on the development of body ownership and agency in **Figure 1** (based on Georgie et al., 2019) and integrate these findings with the development of interoception.

DEVELOPMENT OF INTEROCEPTION AND THE MINIMAL SELF

Interoception refers to perceiving signals from inner organs such as heartbeat, hunger, or breathing (Herbert and Pollatos, 2012). Interoception also includes the monitoring of these internal states during ongoing activities aiming at keeping the bodily system stable (Craig, 2008; Herbert and Pollatos, 2012; Tsakiris, 2017; Seth and Tsakiris, 2018). Before discussing the development of interoception, we would like to note that we differentiate interoceptive sensitivity from interoceptive awareness. Whereas interoceptive sensitivity can be defined as the implicit detection and discrimination of interoceptive signals, interoceptive awareness is a meta-cognitive process reflecting the explicit evaluation of interoceptive states (Murphy et al., 2017). We consider the findings on the development of interoception from this point of view.

Empirical Findings on the Development of Interoception

Similar to research on interoception and its role in the minimal self in adulthood (Herbert and Pollatas, 2012; Tsakiris, 2017; Marshall et al., 2018), research on interoception development has mainly focused on heartbeat perception. In this section, we will first present the developmental changes in the frequency of heartbeats, which will be followed by a review on cardiac interoception in infants, children, and adolescents. An overview of this review can be found in **Figure 1**.

Developmental changes in heartbeat frequency can be divided into four phases: (1) from birth to 1 month of age during which the heart rate increases; (2) from 1 month to 2 years of age, in which the heart rate decreases steeply; (3) from 2 to 6 years, in which the heart rate decreases but less strongly as compared to (4) 6–12 years of age (Fleming et al., 2011). Thus, from birth to childhood up until 12 years of age, pronounced changes occur in the frequency of heartbeats. Similar developmental changes have been documented for cardiac interoceptive abilities (Koch and Pollatos, 2014; Georgiou et al., 2015; Klabunde et al., 2019; Jones et al., 2021).

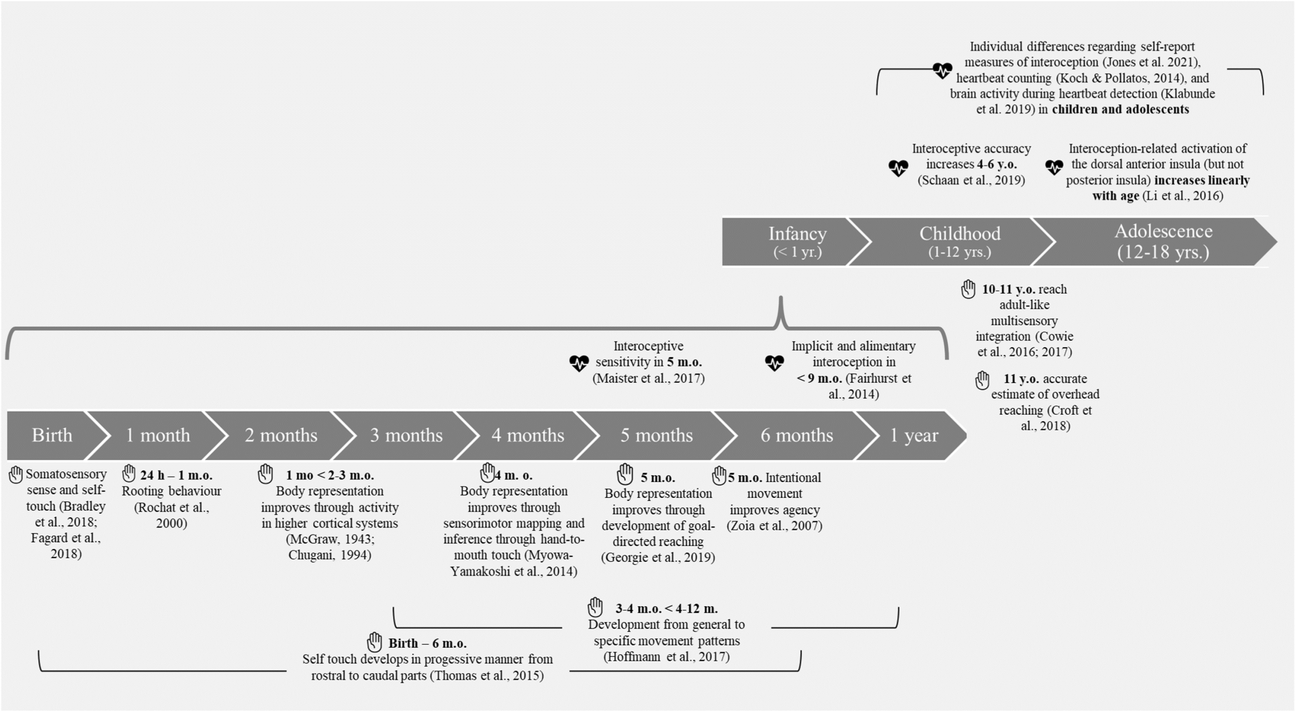

Individual differences regarding self-report measures of interoception (Jones et al. 2021), heartbeat counting (Koch & Pollatos, 2014), and brain activity during heartbeat detection (Klabunde et al. 2019) in **children and adolescents**
Interoceptive accuracy increases **4-6 y.o.** (Schaan et al., 2019)
Interoception-related activation of the dorsal anterior insula (but not posterior insula) **increases linearly with age** (Li et al., 2016)
Infancy (< 1 yr.)
Childhood (1-12 yrs.)
Adolescence (12-18 yrs.)
10-11 y.o. reach adult-like multisensory integration (Cowie et al., 2016; 2017)
11 y.o. accurate estimate of overhead reaching (Croft et al., 2018)
Interoceptive sensitivity in **5 m.o.** (Maister et al., 2017)
Implicit and alimentary interoception in **< 9 m.o.** (Fairhurst et al., 2014)
Birth
1 month
2 months
3 months
4 months
5 months
6 months
1 year
Somatosensory sense and self-touch (Bradley et al., 2018; Fagard et al., 2018)
24 h – 1 m.o. Rooting behaviour (Rochat et al., 2000)
1 mo < 2-3 m.o. Body representation improves through activity in higher cortical systems (McGraw, 1943; Chugani, 1994)
4 m. o. Body representation improves through sensorimotor mapping and inference through hand-to-mouth touch (Myowa-Yamakoshi et al., 2014)
5 m.o. Body representation improves through development of goal-directed reaching (Georgie et al., 2019)
5 m.o. Intentional movement improves agency (Zoia et al., 2007)
3-4 m.o. < 4-12 m. Development from general to specific movement patterns (Hoffmann et al., 2017)
Birth – 6 m.o. Self touch develops in progessive manner from rostral to caudal parts (Thomas et al., 2015)

FIGURE 1 | Overview of studies on the development of interoception as well as body representation, multisensory integration, ownership, and agency relevant for the human minimal self during infancy, childhood, and adolescence. The hand symbol represents studies on body representation, multisensory integration, ownership, and agency. The heart symbol represents studies on interoception. The lower part of the figure summarizes the results of infant studies and the upper part of the figure summarizes the study results on children and adolescents. m.o., month-olds; y.o., year-olds.

In infancy (up to 1 year) and early childhood (1–5 years), very few empirical studies investigated interoceptive abilities (Fairhurst et al., 2014; Maister et al., 2017). The only published empirical study investigating cardiac interoception in infants suggests that, already by 5 months of age, infants show sensitivity to their cardiac signals (Maister et al., 2017). In this study, infants were presented with images that moved synchronously or asynchronously with their own heartbeat. Infants looked significantly longer at asynchronously presented stimuli suggesting that they were able to distinguish asynchronous from synchronous stimuli (Maister et al., 2017). Moreover, individual differences in looking times were correlated with heartbeat-evoked potentials, a brain signal related to cardiac interoceptive processing (Coll et al., 2021). In other words, infants who responded to the synchronous manipulation also showed stronger neural responses captured by the heartbeat-evoked potentials. These findings support the argument that interoception may contribute to the development of the minimal self.

In children (>5 to 12 years) and adolescents (12–18 years), interoception has been investigated mostly by adopting approaches and methodologies used in adult populations. Empirical findings suggest that, similar to adults, children and adolescents show individual differences in heartbeat counting tasks (Koch and Pollatos, 2014) and self-report measures of interoception such as those collected through the Multidimensional Assessment of Interoceptive Awareness Questionnaire (Jones et al., 2021). By inducing cardiac perturbation through jumping jacks and assessing heartbeat counting abilities before and after the tasks, researchers have shown that children accurately count their heartbeats as early as 4–6 years of age (Schaan et al., 2019). Moreover, brain areas such as the left insula, cuneus, inferior parietal lobule, and prefrontal regions are activated during a heartbeat detection task already at 6 years of age (Klabunde et al., 2019).

Studies in children and adolescents also indicated age-related differences in interoception. For example, children's performance in an adapted version of the heartbeat counting task increases with age, which marginally predicts emotion regulation, but not emotion recognition (Koch and Pollatos, 2014). Moreover, during the heartbeat detection task adolescents show increased activation in brain regions related to meta-cognition such as the dorsal anterior cingulate cortex, orbital frontal cortex, and mid-inferior frontal gyrus as compared to children (Klabunde et al., 2019). This neural pattern of activation might suggest that meta-cognitive aspects of interoceptive processing might develop throughout adolescence.

Overall, the empirical results describing the developmental trajectory of interoception in childhood, and especially in infancy, are scarce but much needed. Among others, this scarcity of research is likely due to methodological challenges in measuring interoception in younger children. Next, we extend the existing embodiment account on interoception and formulate theoretical predictions on the development of interoception for future research.

Theoretical Predictions

In the following, we formulate developmental predictions derived from an embodied cognition account of interoceptive inference. Importantly, our embodied cognition perspective assumes that representations form the body-goal link (cf., Pacherie, 2018; Raab and Araujo, 2019; see Witt and Riley, 2014 for alternative accounts considering interoception), enable goal-directed acting in a flexible and adaptive manner (Schulz, 2018), as well as emerge through sensorimotor and bodily experiences throughout development (cf., Musculus et al., 2021). Given the scarcity of research on the development of interoception, and particularly on interoceptive modalities such as respiration, thermoregulation and so forth, we center our arguments on cardiac interoception from birth to 12 years of age. We focus on this age range based on (1) the developmental changes in the frequency of heartbeats (Fleming et al., 2011), (2) motor and bodily development (Musculus et al., 2021), and (3) findings on multisensory integration, the sense of ownership and agency (see **Figure 1**; cf., Georgie et al., 2019). We point out the interaction between multisensory integration of external sensory input, ownership, and agency with internal bodily signals in the formation of the minimal self in development.

Interoceptive sensitivity is observed in the first months of life (Maister et al., 2017). Interestingly, changes in interoceptive sensitivity coincide with the improvements in sensorimotor mapping such as hand-to-mouth touch (Myowa-Yamakoshi and Takeshita, 2006) and goal-directed reaching (Georgie et al., 2019). Together, these developments might contribute to the formation of body representation, and hence, to the sense of ownership in infants at 5–6 months of age (see **Figure 1**). Through improvements in motor skills and continuous exploration, infants learn to act in a goal-directed manner (i.e., goal-directed touching and reaching). This, in turn, helps them to learn about their body boundaries and relate body-directed goals (e.g., reaching the mouth) to goals in the environment. Establishing this relation might pave the way to a sense of body ownership in humans.

Moving further in the developmental trajectory, we hypothesize that the first 2 years of life are crucial to study the development of interoception. This prediction is based on the rapid decrease in heart-beat frequency until 2 years of age (Fleming et al., 2011) and the rather general developmental embodied cognition premise that phases of rapid bodily changes and motor development promote perceptual and cognitive changes (Loeffler et al., 2016; Musculus et al., 2021). We further hypothesize that there might be more drastic changes in interoceptive sensitivity between 2 and 6 years of age (i.e., phases of rapid growth and motor learning) as compared to 6–12 years of age. Moreover, we expect interoceptive awareness to develop during late childhood to adolescence. This change is likely due to the development of meta-cognitive processes (Klabunde et al., 2019). The developmental changes in interoception coincide with improvements in multisensory integration (Cowie et al., 2016, 2017) and accuracy of reach estimations (Croft et al., 2018), which might indicate more accurate representation of the body–environment relation. This relation might be further mediated by an increase in confidence in judging bodily as well as motor competences.

Further, we specify the relationship between interoception and other minimal-self dimensions such as the sense of ownership and agency. To do so, we dissociate a low-level agency (i.e., agency feeling) from a high-level agency (i.e., agency judgment; Synofzik et al., 2008). We assume that this distinction develops with age. First, we hypothesize that interoceptive sensitivity and body ownership are functionally and reciprocally interconnected. That is, improvements in perceiving and identifying internal bodily signals (i.e., interoception) as well as the boundary between one's body and the external environment (i.e., body ownership) should benefit one another. For example, perceiving one's heartbeat might promote the feeling of the body as one's own. Moreover, we hypothesize that improvements in interoceptive awareness in late childhood or adolescence could coincide with a high-level agency judgment due to the involvement of meta-cognitive processes. Overall, we argue that considering the interaction between interoception, other minimal-self components and bodily development is crucial to define, test, and disentangle mechanisms underlying minimal-self development.

Future Research and Conclusion

We suggest a research program to empirically test the predictions on the development of interoception. The program entails specific study designs and a psychophysiological multi-method approach to capture the developmental trajectory of interoception as well as its relation to other minimal-self components such as ownership and agency.

We need longitudinal designs to test the developmental trajectories. Longitudinal designs enable us to disentangle intraindividual changes over the course of development as well as interindividual differences when people of the same age develop differently. Moreover, training studies would inform our understanding of the relationship between bodily changes and interoception. In training studies different training groups differentially targeting the bodily system could be implemented to look at the respective effects on interoception. For instance, infants and children could engage in physical exercises that either lead to an increase or a decrease in their heart rate and the respective effects on interoceptive abilities could be measured.

To investigate the link between interoception and other minimal-self components such as ownership and agency, measurements from both lines of research need to be combined. Therefore, we suggest that interoception paradigms should be jointly implemented with body representation (cf., Suzuki et al., 2013) and multisensory integration paradigms in infant and child studies (e.g., Cowie et al., 2016, 2017). Studies combining measures within the same developmental study would improve our understanding of how multiple sources of bodily and sensory information contribute to the development of the self. This would allow us to better understand how ownership and agency relate to and change in relation to interoception.

In combination, developmental study designs and a psychophysiological multi-method approach (Hoffmann et al., 2018) could even help testing potentially competing mechanisms (Marshall et al., 2018; Seth and Tsakiris, 2018) on the relation between interoception and (motor) action and their respective contribution to the minimal self. Combining cohort-longitudinal designs by enrolling infants and children of different ages with simultaneously applying cardiac-physiological (electrocardiography), neural (electroencephalography), and motor (electromyography) measures might help disentangle these mechanisms. In particular, event-related, reaction-time paradigms could be used that require a motor response. At the same time cardiac and motor measures could be combined to infer how interoceptive and motor states functionally interact in the same experimental task.

There are other developmental aspects that we do not elaborate on due to our focus on childhood rather than infancy. However, we deem the following aspects relevant for future work on interoception: The relation between interoception and active self-touch as well as the role of social interactions. Infancy work has lately also considered the link between interoception and haptic perception (i.e., active self-touch; Fotopoulou and Tsakiris, 2017). This work suggests that active self-touch might benefit the later integration of tactile-proprioceptive and visual information relevant for minimal-self development (see Nguyen et al., 2021 for a review). Besides, social interactions have been considered to play a crucial role in the development of the minimal self, particularly in the development of interoceptive abilities in early infancy (Fotopoulou and Tsakiris, 2017). Given that infants are born with limited motor skills, they depend on others to regulate their own bodily needs such as hunger. Thus, infants rely on embodied interactions with their caregivers in order to regulate their interoceptive states. These interactions allow them to learn the regularities within and outside their bodies (Tsakiris, 2017). Future studies should empirically test the role of embodied interactions in the construction of the minimal self early on in life, including all aspects such as interoception, agency, and ownership.

To sum up, a comprehensive research program is warranted. Such a program would further benefit from a new psychophysiological approach (Hoffmann et al., 2018) and from studying social aspects of interoception (Fotopoulou and Tsakiris, 2017). Together, we hope that the theoretical predictions and the research program introduced in this perspective will promote future research to understand the role of interoception in the development of the minimal self.

AUTHOR CONTRIBUTIONS

LM, MT, MR, and EK contributed to the conceptualization and wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

LM and MR were funded by the German Research Foundation (DFG; RA 940/21-1). EK was funded by the DFG (grant number: 402789467) and the Open Access Publishing Fund of University of Potsdam. The funders had no role in conceptualization, and development or writing of the manuscript. MT was funded by the Austrian Science Fund (FWF), project number 33486-B.

REFERENCES

Chemero, A. (2011). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press/Bradford Books.

Ciaunica, A., Constant, A., Preissl, H., and Fotopoulou, K. (2021). The first prior: from co-embodiment to co-homeostasis in early life. *Conscious. Cogn.* 91:103117. doi: 10.1016/j.concog.2021.103117

Coll, M. P., Hobson, H., Bird, G., and Murphy, J. (2021). Systematic review and meta-analysis of the relationship between the heartbeat-evoked potential and interoception. *Neurosci. Biobehav. Rev.* 122, 190–200. doi: 10.1016/j.neubiorev.2020.12.012

Cowie, D., McKenna, A., Bremner, A. J., and Aspell, J. E. (2017). The development of bodily self-consciousness: changing responses to the full body illusion in childhood. *Dev. Sci.* 21, 1–12. doi: 10.1111/desc.12557

Cowie, D., Sterling, S., and Bremner, A. J. (2016). The development of multisensory body. Representation and awareness continues to 10 years of age: evidence from the rubber hand illusion. *J. Exp. Child Psychol.* 142, 230–238. doi: 10.1016/j.jecp.2015.10.003

Craig, A. D. (2008). "Interoception and emotion: a neuroanatomical perspective," in *Handbook of Emotions. Vol. 3.* eds. M. Lewis, J. M. Haviland-Jones and L. F. Barrett (New York: The Guilford Press), 272–288.

Croft, J. L., Pepping, G. J., Button, C., and Chow, J. Y. (2018). Children's perception of action boundaries and how it affects their climbing behavior. *J. Exp. Child Psychol.* 166, 134–146. doi: 10.1016/j.jecp.2017.07.012

Fairhurst, M. T., Löken, L., and Grossmann, T. (2014). Physiological and behavioral responses reveal 9-month-old infants' sensitivity to pleasant touch. *Psychol. Sci.* 25, 1124–1131. doi: 10.1177/0956797614527114

Fleming, S., Thompson, M., Stevens, R., Heneghan, C., Plüddemann, A., Maconochie, I., et al. (2011). Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *Lancet* 377, 1011–1018. doi: 10.1016/S0140-6736(10)62226-X

Fotopoulou, A., and Tsakiris, M. (2017). Mentalizing homeostasis: the social origins of interoceptive inference. *Neuropsychoanalysis* 19, 3–28. doi: 10.1080/15294145.2017.1294031

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Georgie, Y. K., Schillaci, G., and Hafner, V. V. (2019). "An interdisciplinary overview of developmental indices and behavioral measures of the minimal self." in *9th International Conference on Development and Learning and Epigenetic Robotics;* August 19–22, 2019; Oslo, Norway (IEEE), 129–136.

Georgiou, E., Matthias, E., Kobel, S., Kettner, S., Dreyhaupt, J., Steinacker, J. M., et al. (2015). Interaction of physical activity and interoception in children. *Front. Psychol.* 6:502. doi: 10.3389/fpsyg.2015.00502

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Rev. Philos. Psychol.* 3, 71–88. doi: 10.1007/s13164-012-0089-0

Hafner, V., Hommel, B., Kayhan, E., Lee, D., Paulus, M., and Verschoor, S. (2021). The mechanisms underlying the human minimal self. Available at: https://www.frontiersin.org/research-topics/18147/the-mechanisms-underlying-the-human-minimal-self (Accessed September 23, 2021).

Hafner, V., Hommel, B., and Paulus, M. (2017). The active self (excerpt from the proposal). Available at: http://activeself.de/wp-content/uploads/2017/05/Proposal-SPP-Self-2017-public.pdf (Accessed July 28, 2021).

Herbert, B. M., and Pollatos, O. (2012). The body in the mind: on the relationship between interoception and embodiment. *Top. Cogn. Sci.* 4, 692–704. doi: 10.1111/j.1756-8765.2012.01189.x

Hoffmann, S., Borges, U., Bröker, L., Laborde, S., Liepelt, R., Lobinger, B. H., et al. (2018). The psychophysiology of action: a multidisciplinary endeavor for integrating action and cognition. *Front. Psychol.* 9:1423. doi: 10.3389/fpsyg.2018.01423

Jacob, P. (2016). "Assessing radical embodiment," in *Foundations of Embodied Cognition—Volume 1. Perceptual and Emotional Embodiment*. eds. M. H. Fischer and Y. Coello (London, England: Taylor and Francis), 38–58.

Jones, A., Silas, J., Todd, J., Stewart, A., Acree, M., Coulson, M., et al. (2021). Exploring the multidimensional assessment of interoceptive awareness in youth aged 7–17 years. *J. Clin. Psychol.* 77, 661–682. doi: 10.1002/jclp.23067

Klabunde, M., Juszczak, H., Jordan, T., Baker, J. M., Bruno, J., Carrion, V., et al. (2019). Functional neuroanatomy of interoceptive processing in children and adolescents: a pilot study. *Sci. Rep.* 9:16184. doi: 10.1038/s41598-019-52776-4

Koch, A., and Pollatos, O. (2014). Cardiac sensitivity in children: sex differences and its relationship to parameters of emotional processing. *Psychophysiology* 51, 932–941. doi: 10.1111/psyp.12233

Loeffler, J., Raab, M., and Cañal-Bruland, R. (2016). A lifespan perspective on embodied cognition. *Front. Psychol.* 7:845. doi: 10.3389/fpsyg.2016.00845

Maister, L., Tang, T., and Tsakiris, M. (2017). Neurobehavioral evidence of interoceptive sensitivity in early infancy. *eLife* 6:e25318. doi: 10.7554/eLife.25318

Marshall, A. C., Gentsch, A., and Schütz-Bosbach, S. (2018). The interaction between interoceptive and action states within a framework of predictive coding. *Front. Psychol.* 9:180. doi: 10.3389/fpsyg.2018.00180

Murphy, J., Brewer, R., Catmur, C., and Bird, G. (2017). Interoception and psychopathology: a developmental neuroscience perspective. *Dev. Cogn. Neurosci.* 23, 45–56. doi: 10.1016/j.dcn.2016.12.006

Musculus, L., Ruggeri, A., and Raab, M. (2021). Movement matters! Understanding the developmental trajectory of embodied planning. *Front. Psychol.* 12:633100. doi: 10.3389/fpsyg.2021.633100

Myowa-Yamakoshi, M., and Takeshita, H. (2006). Do human fetuses anticipate self-oriented actions? A study by four-dimensional (4D) ultrasonography. *Infancy* 10, 289–301. doi: 10.1207/s15327078in1003_5

Newen, A., De Bruin, L., and Gallagher, S. (eds.) (2018). *The Oxford Handbook of 4E Cognition*. Oxford, England: Oxford University Press.

Nguyen, P. D. H., Georgie, Y. K., Kayhan, E., Eppe, M., Hafner, V. V., and Wermter, S. (2021). Sensorimotor representation learning for an "active self" in robots: a model survey. *Künstl. Intell.* 35, 9–35. doi: 10.1007/s13218-021-00703-z

Pacherie, E. (2018). "Motor intentionality," in *The Oxford Handbook of 4E Cognition*. eds. A. Newen, L. De Bruin and S. Gallagher (Oxford, UK: Oxford Press), 369–387.

Raab, M., and Araújo, D. (2019). Embodied cognition with and without mental representations: the case of embodied choices in sports. *Front. Psychol.* 10:1825. doi: 10.3389/fpsyg.2019.01825

Schaan, L., Schulz, A., Nuraydin, S., Bergert, C., Hilger, A., Rach, H., et al. (2019). Interoceptive accuracy, emotion recognition, and emotion regulation in preschool children. *Int. J. Psychophysiol.* 138, 47–56. doi: 10.1016/j.ijpsycho.2019.02.001

Schulz, A. W. (2018). *Efficient Cognition: The Evolution of Representational Decision Making*. Cambridge, MA: MIT Press.

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

Seth, A. K., and Tsakiris, M. (2018). Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* 22, 969–981. doi: 10.1016/j.tics.2018.08.008

Shapiro, L. (2019). *Embodied Cognition. 2nd Edn.* London: Routledge/Taylor & Francis Group.

Sterling, P. (2014). Homeostasis vs allostasis: implications for brain function and mental disorders. *JAMA Psychiat.* 71, 1192–1193. doi: 10.1001/jamapsychiatry.2014.1043

Suzuki, K., Garfinkel, S. N., Critchley, H. D., and Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia* 51, 2909–2917. doi: 10.1016/j.neuropsychologia.2013.08.014

Synofzik, M., Vosgerau, G., and Newen, A. (2008). I move, therefore I am: a new theoretical framework to investigate agency and ownership. *Conscious. Cogn.* 17, 411–424. doi: 10.1016/j.concog.2008.03.008

Tsakiris, M. (2017). The multisensory basis of the self: from body to identity to others. *Q. J. Exp. Psychol.* 70, 597–609. doi: 10.1080/17470218.2016.1181768

Tsakiris, M., Jiménez, A. T., and Costantini, M. (2011). Just a heartbeat away from one's body: interoceptive sensitivity predicts malleability of body-representations. *Proc. R. Soc. B Biol. Sci.* 278, 2470–2476. doi: 10.1098/rspb.2010.2547

Verschoor, S. A., and Hommel, B. (2017). Self-by-doing: the role of action for self-acquisition. *Soc. Cogn.* 35, 127–145. doi: 10.1521/soco.2017.35.2.127

Wilson, M. (2002). Six views of embodied cognition. *Psychon. Bull. Rev.* 9, 625–636. doi: 10.3758/bf03196322

Witt, J. K., and Riley, M. A. (2014). Discovering your inner Gibson: reconciling action-specific and ecological approaches to perception-action. *Psychon. Bull. Rev.* 21, 1353–1370. doi: 10.3758/s13423-014-0623-4

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

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

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 11 November 2021
doi: 10.3389/fpsyg.2021.712559

Environment-Related and Body-Related Components of the Minimal Self

Marvin Liesner and Wilfried Kunde*

Department of Cognitive Psychology, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

Perceptual changes that an agent produces by efferent activity can become part of the agent's minimal self. Yet, in human agents, efferent activities produce perceptual changes in various sensory modalities and in various temporal and spatial proximities. Some of these changes occur at the "biological" body, and they are to some extent conveyed by "private" sensory signals, whereas other changes occur in the environment of that biological body and are conveyed by "public" sensory signals. We discuss commonalties and differences of these signals for generating selfhood. We argue that despite considerable functional overlap of these sensory signals in generating self-experience, there are reasons to tell them apart in theorizing and empirical research about development of the self.

Keywords: active self, exteroception, ideomotor theory, interoception, minimal self, self-construction, sense of agency, sense of ownership

OPEN ACCESS

Edited by:
Ezgi Kayhan,
University of Potsdam, Germany

Reviewed by:
Niclas Braun,
University Hospital Bonn, Germany
Michel Cermolacce,
Assistance Publique, France

****Correspondence:***
Marvin Liesner
marvin.liesner@uni-wuerzburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *20 May 2021*
Accepted: *12 October 2021*
Published: *11 November 2021*

Citation:
Liesner M and Kunde W (2021) Environment-Related and Body-Related Components of the Minimal Self. Front. Psychol. 12:712559. doi: 10.3389/fpsyg.2021.712559

INTRODUCTION

Which type of systems, biological or artificial, might develop a self? According to sensorimotor approaches of the self, only agents can do so (Gallagher, 2000; Verschoor and Hommel, 2017). Agents are systems that process sensory data and generate efferent activity, which changes these sensory data. In other words, agents are systems that act and perceive. A developing human can become an agent, so as a (simulated) robot can possibly do (Hafner et al., 2020). In humans, perception relates to all kinds of sensory data, which come across as different perceptual modalities (like vision, audition, proprioception, etc.), while efferent activity is generated by muscle contractions. Robots perceive and act depending on their sensory and motor equipment.

Agents develop a "minimal self" (Gallagher, 2000), provided two learning processes take place. First, agents sense that they causally change perceptual states by efferent activity. In humans, this perceived causality between one's voluntary action and a perceived outcome is called a "sense of agency," which can be measured in various ways (Haggard, 2017). For a robot to develop a minimal self, a corresponding representation of this causal knowledge is required (Hafner et al., 2020). Second, the agent has to sense that there is a spatially extended part of the perceptual world that is somehow "unique" in that it certainly identifies the agent as the source of a perceptual experience. This unique part of the world is the agent's "body" (Gallagher, 2000). In humans, this experience is called sense of body "ownership," and like sense of agency, it can be assessed in various ways including explicit and implicit measures (Tsakiris, 2010, 2017). In robots, the development of a corresponding representation of the physical extension of the robot might be construed as a sense of body ownership as well (Hoffmann et al., 2018; Zenha et al., 2018).

In this article, we discuss reasons to tell apart two types of perceptual events, body-related and environment-related (exteroceptive) events. Moreover, we argue that body-related events should be ascribed a special role when it comes to develop a minimal self. Body-related events include, but are not restricted to, sensory events, which are often subsumed under the term interoception. Interoception is often meant to describe various kinds of sensory signals that originate from biological bodies, including visceral signals such as heart rate and body temperature (Craig, 2009; Tsakiris, 2017). While we acknowledge the role of such visceral signals for self-development (cf., Gentsch and Schütz-Bosbach, 2015; Marshall et al., 2018), we use the term interoception to refer to body-related signals that originate from moving the body, such as proprioception and tactile perception. We are aware that this does not quite match the common definition of interoception. However, in lack of a better and more specific term to subsume proprioception and tactile perception and to avoid the clumsiness and missing precision of speaking of "body-related signals" throughout the manuscript, we will use the term interoception in the following to summarize these sensory signals. Exteroceptive perception on the other hand relates to sensory processing that has the potential to capture events that are distinct from the agent, such as vision or audition.

Why is it worth discussing, or better reminding of, reasons to distinguish between these interoceptive and exteroceptive events? We think there are two reasons for doing so. First, recent sensorimotor approaches to the self tend to treat them as more or less equivalent (Ma and Hommel, 2015; Verschoor and Hommel, 2017). According to sensorimotor approaches, control over perceptual events is sufficient to integrate these events into the self in terms of sense of agency, and subsequently, the sense of body ownership. For example, Ma and Hommel (2015) conclude "people perceive as their body everything that expresses their intentions, including things within reach that move 'as they wish'" (p. 85). While we are generally sympathetic to this view, we want to highlight that control over exteroceptive events is important, but not sufficient, to induce body ownership experience. To experience ownership of exteroceptive events, control over these events must be accompanied by concurrent control over interoceptive events. Even with such concurrent control, a sense of body ownership of exteroceptive events sometimes fails to occur. For example, controlling an external object like a rubber hand does not come with an ownership experience of that rubber hand if it is placed in an anatomically implausible position (Kalckert and Ehrsson, 2012). Also, controlling a tool does not come with an ownership experience of the tool if it moves spatially incompatible to the operating hand (Liesner et al., 2020b). Second, to consider the role of interoceptive events might be particularly relevant when it comes to model how artificial agents, like robots, may or may not develop a self. If there is a special role in interoceptive events in human agents when it comes to self-development, as we believe, this raises the question in which way robots develop a self, similar to that humans have. Or, put differently: If one has the aim to develop a self in robots with sufficient similarity to the human self, how can one then account for this special role of interoceptive events? We want to make our case by assessing the role of interoceptive and exteroceptive events for two key components of a minimal self, the sense of agency and the sense of ownership. We then conclude by discussing some avenues for future research.

SENSE OF AGENCY

What does it take to develop a sense of agency? Empirical research has shown that a match is important between perceptual changes the agent aimed at prior to generating efferent activity (often called "goals"), and the actual perceptual feedback after the efferent activity had been emitted (for a recent overview, refer Haggard, 2017). If there is a match between goal and feedback, it is likely that this feedback was caused by the agent's efferent activity (Carruthers, 2012; Haggard and Chambon, 2012; Gallagher, 2013; Zaadnoordijk et al., 2019). If there is a mismatch, it is more likely that the postaction percept was caused by something else than efferent activity. In case of repeated matches between anticipation and perceptual feedback, this feedback can said to be controllable.

Models of motor control vary regarding the functional role that anticipation of feedback has. Some models construe such anticipations as "predictions" and the corresponding mismatch with actual perceptual feedback as a "prediction error" (Miall and Wolpert, 1996; Wolpert, 1997; Wolpert and Flanagan, 2001). These models thus assume that predictions are derived from the already specified efferent activity. We tend to favor a different view, which is called ideomotor control (James, 1981; Koch et al., 2004; Shin et al., 2010; Waszak et al., 2012; Hommel, 2013). This model assumes that the agents first accidentally produce efferent activity (motor babbling) and link the consistently ensuing perceptual changes to that efferent activity. Only after such links have been established, can an efferent activity be deliberately generated by recollecting the consistently produced perceptual changes, which are then named "goals." Briefly, efferent activity is accessed through perceptual goals.

In this ideomotor model, prediction does not have a strong role. The perceptual goal is the best prediction that agents can possibly have about the outcome of their efferent activity, which is the very reason for activating this specific motor activity. Without these perceptual goals, they could not move intentionally at all. There is ample evidence suggesting that motor activities are indeed generated by recollecting their associated, and currently intended, perceptual changes (e.g., Elsner and Hommel, 2001; Kunde, 2001; Liesner et al., 2020a). To illustrate the difference between the two approaches, consider an example of a simple grasping action: Prediction-based models would assume that having the intention to achieve a certain end state of the action (i.e., grasping the object) triggers the implementation of a motor plan to achieve this intention. Based on this motor plan, a perceptual prediction is derived of how it should "feel" to achieve the intended end state of grasping the object, a so-called "efference copy." The actual sensation while grasping is then compared with this predicted state (Miall and Wolpert, 1996; Wolpert, 1997; Wolpert and Flanagan, 2001). Ideomotor models, however, assume that the intended end state is essentially already an anticipation of the sensory consequences of the action.

According to ideomotor models, one would thus anticipate how it "feels" to grasp the object and this anticipation would then trigger the necessary motor activities to achieve this sensory state. This link between motor activities and sensory effects is based on one's learning history, which specific motor activities (e.g., grasping movements) and sensory effects (e.g., grasping sensations) have frequently occurred together (James, 1981; Koch et al., 2004; Shin et al., 2010; Waszak et al., 2012; Hommel, 2013). A sense of agency would then be inferred from the match between intended effects (i.e., goals) and actually observed effects. We tend to favor this in our view more parsimonious ideomotor approach over the prediction-based approach. We think that adding a further sensory prediction to the action planning phase when the intended sensory state is already known does not provide much benefit for the agent, and seems dispensable to explain agency and ownership experiences. We do not have the space here to discuss possible distinctions between "predictions" and "goals" further. For the purpose of the present paper, a strict differentiation of the two is not necessary since the model we want to propose in this paper only suggests that sensory anticipations of some form are made when engaging in the voluntary efferent activity. Moreover, different views on predictions and goals are actually not that incommensurate. Recent approaches suggest that predictions are the "motor commands" that generate efferent activity according to ideomotor theory (Brown et al., 2013; for a discussion of predictions versus goals see Dogge et al., 2019).

Here comes an important point. Depending on the sensory equipment of the agent, efferent activities typically produce all kinds of sensory feedback. Similarly, agents can have all kinds of perceptual goals. They might generate the same efferent activity at a superficial level to reach these different goals (Kunde and Weigelt, 2005; Pfister, 2019; Mocke et al., 2020). Think of a person controlling a tool such as a mouse cursor on a PC screen. After some experience with the tool (i.e., after associations between muscle contractions and cursor movements have been established), an agent might generate a movement of the tool by recollecting the visual tool trajectory (i.e., anticipating the cursor movements on the screen). Yet, the agent might also produce the superficially same movement by recollecting the proprioceptive sensations of the corresponding hand movement. There is in fact evidence that agents prefer either one or the other type of perceptual goal, depending on certain factors such as the spatial match between visual and proprioceptive feedback of the motor pattern and the specific task demands (Heuer and Rapp, 2012; Liesner and Kunde, 2020). Is there room for a special role of interoceptive (e.g., proprioceptive) compared with exteroceptive (e.g., visual) motor feedback? Not really. Perhaps the only special role of interoception is that, due to lifelong experience, starting before birth, human agents amass conceivably closer links between efferent activities and interoceptive feedback than they do with any possible exteroceptive feedback. But that is just a gradual rather than a qualitative difference.

But does this mean that every controllable perceptual state becomes part of the self, so as sensorimotor approaches to the self suggest (Ma and Hommel, 2015; Verschoor and Hommel, 2017)? There are both empirical findings and logical arguments that suggest that this is not the case. For example, studies investigating different measures of the sense of agency found that participants experienced less agency when efferent activities led to spatially discrepant interoceptive and exteroceptive signals than when there was no such discrepancy, despite equal controllability (Ebert and Wegner, 2010; Liesner et al., 2020a). According to ideomotor theory, this effect is due to the links of discrepant signals with different, conflicting motor patterns. Most importantly, however, these results are only explainable when keeping up a conceptual differentiation between interoceptive and exteroceptive effects of efferent activities. It has been shown that similar discrepancies between exteroceptive effects only do not lead to such a reduction in the sense of agency (Grechuta et al., 2019). Furthermore, if it would just be controllability of sensory input that determines what we call self, essentially everything we see was part of our self: If we move the eyes to the left, everything on the retina moves to the left. Therefore, every visual object a human can perceptually manipulate by moving the eyes (essentially every visual object) would be part of the self. While this motor-sensory contingency is for sure important to develop consciousness (O'Regan and Noë, 2001; O'Regan, 2011), not every stimulation that reaches consciousness is construed as being part of the self. Also, if it would just be controllability of perceived objects, which determines inclusion of these objects to the self, an agent could not tell apart a mirror image of the agent from the agent. This is sometimes portrayed in a slightly simplified manner in research of self-development in robots. A robot might well detect that it controls a visual mirror image (Hoffmann et al., 2021), but that does not mean that it has developed a self. By contrast, human agents and many animals, starting from a certain age on, can distinguish their "body" from a mirror image of their body (e.g., Gallup, 1970; Amsterdam, 1972; Reiss and Marino, 2001). But how can they do so?

SENSE OF BODY OWNERSHIP

In humans, and perhaps other biological agents, the likely answer to this question is: Because there are unique perceptual events, processed by specific neuronal pathways and cortical regions like the insular, anterior cingulate, or somatosensory cortex (Critchley et al., 2004; Craig, 2009), which can be summarized under the heading interoception. In the context of self-development, the term "interoception" might be a bit misleading, because it suggests that there was already something "interior" (inside the body) and something "exterior" (outside the body), which is the very distinction that the system has to develop in the first place. The crucial point is, however, that there is one, and only one, and thus unique object in the world that can generate "interoceptive" perception, the object that human agents call their "body." For example, we can see that an object touches another object or another agent, so as we can see that an object touches the hand. Yet, only the hand generates the specific perceptual experience of being touched. In psychological theorizing around the concepts of mirroring or empathy, it is sometimes suggested that observers could directly perceive "feelings" or internal states of an observed other agent (e.g., Singer and Lamm, 2009). No, they cannot. The agents might directly see or hear another agent moving, so as they can directly see or hear themselves moving. But only indirectly, by matching that visual or auditory experience to corresponding interoceptive sensations, including those that originate from own moving limbs, the agents might ascribe interoceptive states to another agent (Rizzolatti, 2005; Schütz-Bosbach and Prinz, 2007). Therefore, the agents also cannot mirror the "feeling" of another observed agent, if they have no sufficient recollection of experiencing this "feeling" before themselves (Bosbach et al., 2005). Moreover, the agents cannot imitate other agents, without establishing a linkage between exteroception and interoception through observation of own motor activities. Reports of "imitation" in newborns without that correspondence experience have been criticized on empirical grounds (Slaughter, 2021), or as being expressions of an innate stimulus-response link, where the seen action of a model accidentally matches the innate response of the imitator when judged from a third party (Heyes, 2001). A similar argument has been put forward by phenomenological philosophers in the context of the so-called "analogy argument." This argument suggests that the agents only have access to the internal states of other agents by inferring these from observing the other agents' external states and drawing conclusions based on their own experiences with typical combinations of internal and external states within themselves (Husserl, 1973; Zahavi, 2001). Some authors have even suggested that only because of one's experience with own interoceptive and exteroceptive sensations accompanying each other, one can also understand the existence of others and their selves as entities that are different from one('s)self (Merleau-Ponty, 1945, 1964; Husserl, 1959). Differentiating between interoceptive and exteroceptive signals would thus not only be essential for developing a sense of self, but also for recognizing other agents, which is a crucial skill in the inherently social world that we as humans live in.

It should be noted that the relevant aspect of interoceptive sensory signals for selfhood experiences is not their sensory modalities *per se*, but rather that they diagnostically and infallibly signal the presence of an agent's physical body. In healthy human agents, this function is taken by interoceptive signals, however, in principle, this function could also be taken by other signals, given that they are "exclusive" enough for providing information about the agent's body. We will discuss this possibility further in the context of artificial agents and patients suffering from deafferentation (see next paragraph and section "Agents Without Interoceptive Perception").

Put differently, some perceptual effects of motor activities like visual effects are "public." An agent perceives them more or less, so as other agents do. No doubt, this "publicity" is very important, as it allows matching activities of different agents to each other, and ascribing internal states to other agents, among other things. However, to ascribe uniqueness to an agent's body, controllable sensory events that do arise from just this unique object (the "body") and which are apparent to just the agent, are certainly helpful, if not mandatory. As only the agent has these unique experiences, these experiences might be called "private" (i.e., reserved to the observing agent). In technical systems, these need not necessarily be proprioceptive or tactile events like in humans (if the comparison to human sensory systems makes sense at all). But there has to be some kind of perceptual event that no other object except the agent's physical body can generate. Over the past years, some robotics studies have introduced methods that might be possible candidates for such "private" sensations (Nabeshima et al., 2005; Roncone et al., 2014; Hinz et al., 2018; Hoffmann et al., 2018; Lanillos and Cheng, 2018). For example, information read out from the joint positions of the robot have been suggested as a proxy to proprioceptive sensations (Nabeshima et al., 2005), while pressure sensors in an "artificial skin" on the robot have been used as a source for modeling tactile information (Hinz et al., 2018; Hoffmann et al., 2018). It is beyond the scope of this article to evaluate the adequacy of these approaches and whether they can "substitute" the function of interoceptive sensations in humans. The point that we want to make is that some "private" input of whatsoever form is a necessary prerequisite for the development of an (artificial) self.

The idea that interoceptive signals provide very diagnostic information about the presence of one's (bodily) self has already been put forward by other authors when discussing the principle of "immunity to error through misidentification" (Cassam, 1995; Gallagher, 2013). These authors have suggested that, while sensory information that we would label as exteroceptive (e.g., visual) can be misleading regarding whether it stems from one's own body or not, proprioceptive (i.e., interoceptive) information necessarily signals the presence of one's body since it cannot be perceived for anything or anybody else. As Gallagher (2013) explains, proprioceptive perception can still be erroneous in terms of, for example, the perceived position of a body part (see next paragraph), but there can be no erroneous experience of a perceived proprioceptive signal as not stemming from the own biological body. This view of the innate self-reference of proprioceptive signals is very much compatible with our argumentation that interoceptive signals take a special role regarding the formation of (body) ownership experiences. However, while the previous works mainly focused on the impossibility to misjudge interoceptive signals as not originating from one's own body, we want to make the point that a sense of ownership cannot be experienced at all without any unique sensory experiences, like interoceptive sensations in humans.

As mentioned before, motor activities, at least in neurotypical agents, mostly produce public and private signals at the same time. We can see and feel our hand moving or being touched, and we make a repeated experience that these perceptual events normally coincide in space and time, such that we see and feel a hand moving rightward. Because interoceptive events, like touch, are very diagnostic for body ownership, but have a low spatial accuracy, human agents sometimes misjudge visual events as indicating body ownership, if these visual events temporally coincide with interoceptive percepts despite moderate spatial displacement to these corresponding interoceptive events. This is the functional basis behind the so-called rubber hand illusion and other body-transfer illusions (e.g., Slater et al., 2010; Tajadura-Jiménez et al., 2012; Maselli and Slater, 2013). In the original experiment by Botvinick and Cohen (1998), a rubber hand that is seen to be stroked while the own hand is felt being stroked appears as belonging to the body. Thus, the temporal coincidence of interoceptive and exteroceptive events can create the impression that exteroceptive events belong to the same entity that normally produces interoceptive events, the body. Importantly, this, however, does not contradict the conceptual differentiation between interoceptive and exteroceptive sensations that we want to make in this study. Even if exteroceptive events are integrated with interoceptive events and the source of the latter might thus be experienced as belonging to one's body, this does not mean that the interoceptive and exteroceptive sensations are experienced any differently *per se*. In the rubber hand illusion, the stroking on one's real hand is mislocalized on the artificial hand (e.g., Dummer et al., 2009; Rohde et al., 2011; Kalckert and Ehrsson, 2012). However, this does not qualitatively change the interoceptive, tactile sensation felt by the brushstroke in any way. Similarly, also the exteroceptive, visual sensations from the rubber hand are not experienced qualitatively differently. For example, the rubber hand does not look any different for a participant experiencing the rubber hand illusion from what it looks like without experience of the illusion (Botvinick and Cohen, 1998; Rohde et al., 2011). "Integration" of interoceptive and exteroceptive signals thus does not mean that a new "synthesized" percept is created: Instead, some features of the sensation in one modality are shifted toward features in the other modality, the size and direction of which are influenced by the reliability of the sensory signals (Ernst and Banks, 2002; Tsakiris, 2010, 2017; Blanke, 2012). Even in cases of such integration of interoceptive and exteroceptive sensations, a differentiation between them, like we have suggested in this article, still holds.

A coincidence of interoceptive and exteroceptive signals can also be actively generated by efferent activity, thus when moving a body limb that moves another artificial limb ("active rubber hand illusion," Kalckert and Ehrsson, 2012), or another non-corporeal object (Ma and Hommel, 2015). While such body ownership illusions suggest surprising plasticity of what counts as body, they are constrained to cases where there is concurrent, actively produced, interoceptive stimulation. Recently, it has been shown that the mutual relationship of various exteroceptive feedback signals might shape ownership experience (Grechuta et al., 2019), however, also in this case, task-related, interoceptive signals were still present. We are not aware of cases in which the coincidence of, for example, the visual experience of a moving object and a corresponding auditory event alone create ownership experience for that object even close to the range that occurs when interoceptive stimulation is involved.

In biological agents, there is another reason to ascribe interoception a special role. Put simply, every point in space that generates the feeling of touch can bleed, while only few parts of the visual world can do so (those parts of the anatomical body that are visible). It is thus no wonder that biological agents keep an eye on their body, even when they control tools, that otherwise appear to be "embodied" (Collins et al., 2008). After all, on an even higher, reflective level of representation, the lack of some interoceptive stimulation that tools cannot provide is often the very reason for using tools. For example, we use sticks to broil sausages in a campfire rather than our hands. True, depending on the amount of barbecue experience, and corresponding (sense of) agency over the tool, such a tool might appear as being part of the anatomical body (Maravita and Iriki, 2004; Liesner et al., 2020b), but it cannot produce heat pain, which is why we use it. Conceivably, the experience of tool ownership, despite lack of heat perception, does occur only because the tool movements coincide with other interoceptively (i.e., proprioceptively) sensed movements of the operating hand. Additionally, interoceptive signals are not only important to keep the biological substrate of the agent from harm, but they are also essential for the homeostatic and allostatic regulation of the body and the brain (Sterling, 2004, 2012; Barrett et al., 2016; Burleson and Quigley, 2021).

Coincidence of actively generated interoceptive and exteroceptive stimulation is necessary but not sufficient to assign exteroceptive stimulation bodilyness. Specifically, if any object under an agent's immediate control would be experienced by this agent as belonging to their self, the specific relationship of the interoceptive signals from the body controlling the object and of the exteroceptive signals from the object itself should be negligible. Yet, that relationship does count. The explicit and implicit measures of the sense of ownership are decreased or even eliminated when interoceptive or exteroceptive signals are not sufficiently overlapping in terms of direction, location, or timing (e.g., Samad et al., 2015; Pritchard et al., 2016; Kalckert et al., 2019).

To illustrate this point, consider a recent study by Liesner et al. (2020b). The participants were asked to move a visual cursor on a screen by moving their occluded hands. In one (compatible) condition, the cursor moved to the same extent and in the same direction as the felt hand, whereas it moved to the same extent but to the opposite direction as the felt hand, in another (incompatible) condition. At an objective level, controllability of the cursor was identical in both conditions, that is, it was perfectly foreseeable how the cursor would move when then hand moved in both cases. While there were clear indications of ownership experience in the compatible condition, there was no indication of such ownership experience in the incompatible condition. Why is this so? We conjecture that the agents suppress interoceptive codes of their body movements in the incompatible conditions, as these codes cause interference during action planning (Janczyk and Kunde, 2020), a phenomenon also known as "haptic neglect" (Heuer and Rapp, 2012). Because of this suppression of interoceptive codes, it becomes much harder, or even impossible, to establish the coincidence of exteroceptive and interoceptive codes that is crucial to induce a sense of body ownership for exteroceptive events. Thus, exteroceptive codes are not integrated indiscriminately into the self, but only when they match with sufficiently strong interoceptive codes.

INTERIM SUMMARY AND FUTURE DIRECTIONS

Let us briefly summarize. The agents develop a sense of agency, a key component of a minimal self, based on controllable perceptual feedback of their efferent activity. The perceptual feedback in humans comes in various modalities, but there seems

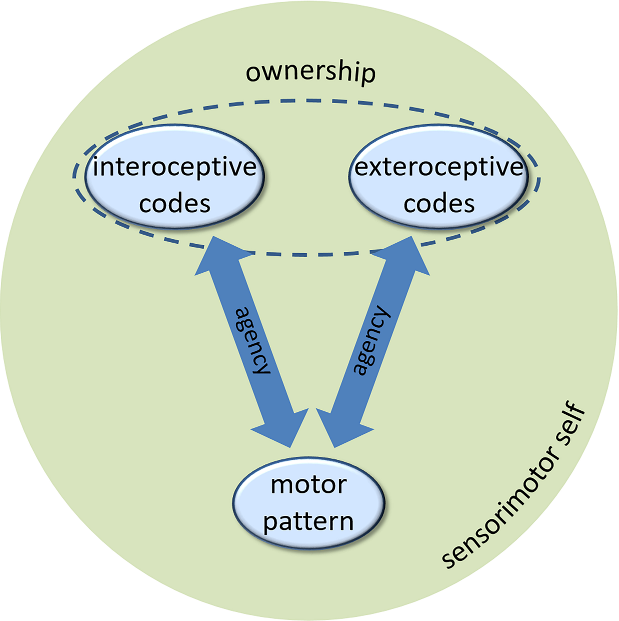


FIGURE 1 | The role of interoceptive and exteroceptive codes in generating the sense of agency and the sense of ownership of a minimal sensorimotor self (see text for description).

no fundamental reason to ascribe interoceptive feedback a special role. The agents can experience agency for interoceptive and exteroceptive events, in the same way, varying, if at all, gradually depending on the strength of associations to the motor patterns that cause these events (cf., **Figure 1**). Yet, there is conceptual and empirical reason to assume that another component of the self, the sense of body ownership, presumes perceptual feedback that no other part of the environment provides. In biological agents, this uniqueness applies to interoceptive sensory signals. In artificial agents, some other conceptualization of this feedback might be possible, given it provides the artificial agent with the same information about the artificial "body" as interoceptive information does for the biological body in biological agents. The sense of ownership of exteroceptive events rests on their integration with interoceptive codes. At the same time, interfering interoceptive and exteroceptive codes of the same action seem to lead to a suppression of the former (Fourneret and Jeannerod, 1998; Knoblich and Kircher, 2004; Müsseler and Sutter, 2009; Sülzenbrück and Heuer, 2009; Heuer and Rapp, 2012; Liesner and Kunde, 2020). While we have demonstrated that ownership experience of exteroceptive events is hard to acquire in these situations, it seems plausible that the unavailability of interoceptive codes either because of haptic neglect or due to loss of neural pathways (as in deafferented patients), might be the causal reason for this. This causal relationship is however, yet to be shown in empirical research. Taking this assessment for granted for a moment, a couple of research questions arise, which we discuss in the following.

Developmental Order of the Sense of Ownership and the Sense of Agency

In the original rubber hand illusion, the ownership experience is induced by "passive" tactile-visual stimulation. This suggests that ownership experience might arise before, or even without, the agent has learned to move in a goal-oriented manner, and hence the experience of agency. However, it is not too far-fetched to assume the exact opposite order of development. As Hinton (2007, p. 535) has put it, "To recognize shapes, first learn to generate images." In other words, to appropriately encode stimulation, observers have to first create that stimulation on purpose. In fact, the interpretation of touch is tightly coupled to active exploration, thus haptics (Lederman and Klatzky, 2009; Bremner and Spence, 2017). Nava et al. (2018) showed that actively applying stroking to oneself in the rubber hand illusion as compared to passively observing an experimenter stroking boosts the illusion in 5-year-old children, while this manipulation is known to rather lead to the opposite effects in adults. Particularly important, and extensively practiced by young infants is double touch, hence, touching the own "body," which creates a tactile experience at both the touching and touched at body part, in the same position in space (Merleau-Ponty, 1954). Perhaps a proper encoding of touch (like being stroked) as diagnostic information for body ownership presumes a sufficient amount of haptic experience, which presumes goal-oriented action, and thus the experience of agency.

Agents Without Interoceptive Perception

If interoceptive perception is the key to derive a sense of ownership, the possible ownership experience of agents without such interoception is a very interesting case. In fact, "deafferented" patients who have lost most of such interoceptive perception, sometimes report having a body that is not that clearly circumscribed. Some report that they experience their body as a tool to affect the environment (Cole and Paillard, 1995). It seems possible that the unique sensory experience of the body that interoceptive perception provides in neurotypical humans becomes substituted by some other (originally exteroceptive) indications of uniqueness, such as the unique visual appearance of the own hands and arms from an ego perspective. However, as the previously mentioned perception of one's body as a tool or other reports of deafferented patients about a disembodied "floating" feeling after the onset of their condition suggests (Cole and Paillard, 1995), this substitution takes time and continued effort to achieve and sustain. These and other alterations in self- and body-perception in deafferented patients (Gallagher and Cole, 1995; Renault et al., 2018) suggest that the innate uniqueness of interoceptive sensations for signaling the presence of one's own body is very difficult, if not impossible, to reach and replace with originally exteroceptive signals.

Deafferented patients are also interesting to study regarding the development of agency experience. In neurotypical human agents, the experience of agency is determined by long-term and short-term links of body movements and visual movement feedback. In most cases throughout lifetime, visual and proprioceptive feedback we get from our body spatially match. If this long-term link is violated by altering visual feedback, such that for example, the visual feedback of a movement is inverted relative to the proprioceptive feedback, as in mirror drawing, agency experience of the visual movement drops (Ebert and Wegner, 2010; Liesner et al., 2020a,b). Moreover, such a violation of long-term sensorimotor experience by short-term alterations of visual feedback comes with considerable drops of performance (Müsseler et al., 2008; Müsseler and Skottke, 2011; Kunde et al., 2012). Interestingly, patients with loss of interoceptive perception do not consistently show such a drop in performance (Lajoie et al., 1992). It seems likely that they do not experience reduced agency either. This may depend, however, on the way the sense of agency is explored. While tactile perception normally shapes the experience of, for example, temporal binding (Haggard, 2017), which is often considered an unobtrusive measure of the sense of agency (Cao et al., 2020), it is conceivable that these patients would still distinguish between normal and mirror drawing in their subjective experience of agency, just like neurotypical agents do (Ebert and Wegner, 2010; Liesner et al., 2020a). However, this agency experience would then most likely be based on the (mis)match of visual feedback in the environment and unique visual body representations, instead of unique proprioceptive body representations. Because many "standard" robots today are not yet equipped with sophisticated "interoceptive" sensors, the study of the sense of ownership and the sense of agency in deafferented patients might be quite inspiring for roboticists who aim to develop machines that contain these cornerstones of selfhood.

Prosthesis Ownership Experiences and Phantom Limbs

Another interesting domain to study the role of intero- and exteroception are patients with limb prosthesis and/or phantom limb experiences. While the former basically provides a situation of a "body" part without any interoceptive sensation, similar to deafferented patients, the latter can be described as a case of illusory interoception (based on previous experiences; Ramachandran, 1998) without a corresponding body part. Recent studies have shown that extended motor control and sensory feedback from using a prosthesis enhances experienced ownership of the prosthesis and reduces phantom limb experiences (Page et al., 2018), while ownership experience of prosthesis and phantom limb experience are negatively correlated (Bekrater-Bodmann et al., 2021). This inverse relationship between prosthesis ownership experience and phantom limb experience might result from the transfer of memories of previous interoceptive perception from the lost limb to the prosthesis. Indeed, both subjective reports of prosthesis users and brain imaging studies suggest that prostheses can phenomenologically and neurally "replace" lost limbs and that the degree to which this happens is related to the level of satisfaction with and acceptance of the prosthesis (Maruishi et al., 2004; Murray, 2004). It might be interesting for future studies and treatment methods to investigate such a causal protective mechanism of prosthesis ownership experiences against often painful phantom limb experiences.

SUMMARY

The "self" is a glamorous term in social sciences. However, boiling down what it takes for an organism to develop a "self" is challenging. Sensorimotor approaches of this problem suggest that perceptual changes that are controllable by efferent activity tend to become part of the self. This approach is fascinating because it suggests that almost every controllable perceptual event in the world, be it visual, auditory, or proprioceptive, can count as self. This is probably true for the experience of agency. Yet, when it comes to developing a sense of having a body (sense of body ownership), there is conceptual and empirical reason to distinguish between proprioceptive or tactile (i.e., interoceptive) events and other controllable perceptual events. Proprioceptive and tactile events are exceptionally diagnostic to determine which parts of the world belong to the agent and which do not, which is of obvious importance to avoid the physical threat to the agent's biological substrate. Lacking control over or perception of such events comes with severe decrements of the body ownership experience. Moreover, while controlled visual or auditory events might as well be construed by the agent as being owned, this happens only when these events coincide in a spatial and temporal manner with corresponding proprioceptive or tactile changes. Given these empirical observations in human agents, constructing machines that lack interoceptive sensation, but still develop a sense of body ownership in a similar manner as humans do, is a challenge.

AUTHOR CONTRIBUTIONS

ML and WK contributed to the conceptualization of the article, literature search, and writing of both the original and re-edited versions of the manuscript. Both authors approved the submitted version.

FUNDING

This research was funded by the German Research Foundation, DFG in the form of a grant awarded to WK within the DFG Priority Program 2134 "The Active Self" (KU 1964/15-1). This publication was also supported by the Open Access Publication Fund of the University of Würzburg. The funders had no role in conceptualization, development, or writing of the manuscript.

ACKNOWLEDGMENTS

This article is part of the Research Topic "The Mechanisms Underlying the Human Minimal Self" in the section Cognition.

REFERENCES

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Dev. Psychobiol.* 5, 297–305. doi: 10.1002/dev.420050403

Barrett, L. F., Quigley, K. S., and Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philos. Trans. R. Soc. B Biol. Sci.* 371:20160011. doi: 10.1098/rstb.2016.0011

Bekrater-Bodmann, R., Reinhard, I., Diers, M., Fuchs, X., and Flor, H. (2021). Relationship of prosthesis ownership and phantom limb pain: results of a survey in 2383 limb amputees. *Pain* 162, 630–640. doi: 10.1097/j.pain.0000000000002063

Blanke, O. (2012). Multisensory brain mechanisms of bodily self-consciousness. *Nat. Rev. Neurosci.* 13, 556–571. doi: 10.1038/nrn3292

Bosbach, S., Cole, J., Prinz, W., and Knoblich, G. (2005). Inferring another's expectation from action: the role of peripheral sensation. *Nat. Neurosci.* 8, 1295–1297. doi: 10.1038/nn1535

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel'touch that eyes see. *Nature* 391, 756–756. doi: 10.1038/35784

Bremner, A. J., and Spence, C. (2017). The development of tactile perception. *Adv. Child Dev. Behav.* 52, 227–268. doi: 10.1016/bs.acdb.2016.12.002

Brown, H., Adams, R. A., Parees, I., Edwards, M., and Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cogn. Process.* 14, 411–427. doi: 10.1007/s10339-013-0571-3

Burleson, M. H., and Quigley, K. S. (2021). Social interoception and social allostasis through touch: legacy of the somatovisceral afference model of emotion. *Soc. Neurosci.* 16, 92–102. doi: 10.1080/17470919.2019.1702095

Cao, L., Steinborn, M., Kunde, W., and Haendel, B. (2020). Action force modulates action binding: evidence for a multisensory information integration explanation. *Exp. Brain Res.* 238, 2019–2029. doi: 10.1007/s00221-020-05861-4

Carruthers, G. (2012). The case for the comparator model as an explanation of the sense of agency and its breakdowns. *Conscious. Cogn.* 21, 30–45. doi: 10.1016/j.concog.2010.08.005

Cassam, Q. (1995). "Introspection and bodily self-ascription," in *The Body and the Self*, eds J. L. Bermudez, A. Marcel, and N. E. Eilan (Cambridge, MA: The MIT Press), 311–336.

Cole, J., and Paillard, J. (1995). "Living without touch and peripheral information about body position and movement: studies with deafferented subjects," in *The Body and the Self*, eds J. L. Bermúdez, A. J. Marcel, and N. Eilan (Cambridge, MA: The MIT Press), 245–266.

Collins, T., Schicke, T., and Röder, B. (2008). Action goal selection and motor planning can be dissociated by tool use. *Cognition* 109, 363–371. doi: 10.1016/j.cognition.2008.10.001

Craig, A. D. (2009). How do you feel – now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nat. Neurosci.* 7, 189–195. doi: 10.1038/nn1176

Dogge, M., Custers, R., and Aarts, H. (2019). Moving forward: on the limits of motor-based forward models. *Trends Cogn. Sci.* 23, 743–753. doi: 10.1016/j.tics.2019.06.008

Dummer, T., Picot-Annand, A., Neal, T., and Moore, C. (2009). Movement and the rubber-hand illusion. *Perception* 38, 271–280. doi: 10.1068/p5921

Ebert, J. P., and Wegner, D. M. (2010). Time warp: authorship shapes the perceived timing of actions and events. *Conscious. Cogn.* 19, 481–489. doi: 10.1016/j.concog.2009.10.002

Elsner, B., and Hommel, B. (2001). Effect anticipation and action control. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 229–240. doi: 10.1037/0096-1523.27.1.229

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Fourneret, P., and Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia* 36, 1133–1140. doi: 10.1016/S0028-3932(98)00006-2

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Gallagher, S. (2013). "First-person perspective and immunity to error through misidentification," in *Consciousness and Subjectivity*, eds S. Miguens and G. Preyer (Berlin: De Gruyter), 245–272.

Gallagher, S., and Cole, J. (1995). Body image and body schema in a deafferented subject. *J. Mind Behav.* 16, 369–389.

Gallup, G. G. (1970). Chimpanzees: self-recognition. *Science* 167, 86–87. doi: 10.1126/science.167.3914.86

Gentsch, A., and Schütz-Bosbach, S. (2015). "Agency and outcome prediction," in *The Sense of Agency*, eds P. Haggard and B. Eitam (New York, NY: Oxford University Press), 217–234.

Grechuta, K., Ulysse, L., Ballester, B. R., and Verschure, P. F. (2019). Self beyond the body: action-driven and task-relevant purely distal cues modulate performance and body ownership. *Front. Hum. Neurosci.* 13:91. doi: 10.3389/fnhum.2019.00091

Hafner, V. V., Loviken, P., Pico Villalpando, A., and Schillaci, G. (2020). Prerequisites for an artificial self. *Front. Neurorobot.* 14:15. doi: 10.3389/fnbot.2020.00005

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 196–207. doi: 10.1038/nrn.2017.14

Haggard, P., and Chambon, V. (2012). Sense of agency. *Curr. Biol.* 22, R390–R392.

Heuer, H., and Rapp, K. (2012). Adaptation to novel visuo-motor transformations: further evidence of functional haptic neglect. *Exp. Brain Res.* 218, 129–140. doi: 10.1007/s00221-012-3013-z

Heyes, C. (2001). Causes and consequences of imitation. *Trends Cogn. Sci.* 5, 253–261. doi: 10.1016/S1364-6613(00)01661-2

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Prog. Brain Res.* 165, 535–547. doi: 10.1016/S0079-6123(06)65034-6

Hinz, N. A., Lanillos, P., Mueller, H., and Cheng, G. (2018). "Drifting perceptual patterns suggest prediction errors fusion rather than hypothesis selection: replicating the rubber-hand illusion on a robot," in *Proceedings of the 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. (Piscataway, NJ: IEEE), 125–132. doi: 10.1109/DEVLRN.2018.8761005

Hoffmann, M., Straka, Z., Farkaš, I., Vavrečka, M., and Metta, G. (2018). Robotic homunculus: learning of artificial skin representation in a humanoid robot motivated by primary somatosensory cortex. *IEEE Trans. Cogn. Dev. Syst.* 10, 163–176. doi: 10.1109/TCDS.2017.2649225

Hoffmann, M., Wang, S., Outrata, V., Alzueta, E., and Lanillos, P. (2021). Robot in the mirror: toward an embodied computational model of mirror self-recognition. *KI Künstliche Intell.* 35, 37–51. doi: 10.1007/s13218-020-00701-7

Hommel, B. (2013). "Ideomotor action control: on the perceptual grounding of voluntary actions and agents," in *Action Science: Foundations of an Emerging Discipline*, eds W. Prinz, M. Beisert, and A. Herwig (Cambridge, MA: MIT Press), 113–136.

Husserl, E. (1959). *Erste Philosophie II (1923–24), Husserliana VIII*. The Hague: Martinus Nijhoff.

Husserl, E. (1973). *Hua I: Cartesianische Meditationen und Pariser Vorträge*. The Hague: Martinus Nijhoff.

James, W. (1981). *The Principles of Psychology*. Cambridge, MA: Harvard University Press. (Original work published 1890).

Janczyk, M., and Kunde, W. (2020). Dual tasking from a goal perspective. *Psychol. Rev.* 127, 1079–1096.

Kalckert, A., and Ehrsson, H. H. (2012). Moving a rubber hand that feels like your own: a dissociation of ownership and agency. *Front. Hum. Neurosci.* 6:40. doi: 10.3389/fnhum.2012.00040

Kalckert, A., Perera, A. T. M., Ganesan, Y., and Tan, E. (2019). Rubber hands in space: the role of distance and relative position in the rubber hand illusion. *Exp. Brain Res.* 237, 1821–1832. doi: 10.1007/s00221-019-05539-6

Knoblich, G., and Kircher, T. T. (2004). Deceiving oneself about being in control: conscious detection of changes in visuomotor coupling. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 657–666. doi: 10.1037/0096-1523.30.4.657

Koch, I., Keller, P., and Prinz, W. (2004). The ideomotor approach to action control: implications for skilled performance. *Int. J. Sport Exerc. Psychol.* 2, 362–375. doi: 10.1080/1612197X.2004.9671751

Kunde, W. (2001). Response-effect compatibility in manual choice reaction tasks. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 387–394. doi: 10.1037/0096-1523.27.2.387

Kunde, W., and Weigelt, M. (2005). Goal congruency in bimanual object manipulation. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 145–156. doi: 10.1037/0096-1523.31.1.145

Kunde, W., Pfister, R., and Janczyk, M. (2012). The locus of tool-transformation costs. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 703–714.

Lajoie, Y., Paillard, J., Teasdale, N., Bard, C., Fleury, M., Forget, R., et al. (1992). Mirror drawing in a deafferented patient and normal subjects: visuoproprioceptive conflict. *Neurology* 42, 1104–1104. doi: 10.1212/WNL.42.5.1104

Lanillos, P., and Cheng, G. (2018). "Adaptive robot body learning and estimation through predictive coding," in *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. (Piscataway, NJ: IEEE), 4083–4090. doi: 10.1109/IROS.2018.8593684

Lederman, S. J., and Klatzky, R. L. (2009). Haptic perception: a tutorial. *Atten. Percept. Psychophys.* 71, 1439–1459. doi: 10.3758/APP.71.7.1439

Liesner, M., and Kunde, W. (2020). Suppression of mutually incompatible proprioceptive and visual action effects in tool use. *PLoS One* 15:e0242327. doi: 10.1371/journal.pone.0242327

Liesner, M., Kirsch, W., Pfister, R., and Kunde, W. (2020b). Spatial action–effect binding depends on type of action–effect transformation. *Attent. Percept. Psychophys.* 82, 2531–2543. doi: 10.3758/s13414-020-02013-2

Liesner, M., Kirsch, W., and Kunde, W. (2020a). The interplay of predictive and postdictive components of experienced selfhood. *Conscious. Cogn.* 77:102850. doi: 10.1016/j.concog.2019.102850

Ma, K., and Hommel, B. (2015). Body-ownership for actively operated non-corporeal objects. *Conscious. Cogn.* 36, 75–86. doi: 10.1016/j.concog.2015.06.003

Maravita, A., and Iriki, A. (2004). Tools for the body (schema). *Trends Cogn. Sci.* 8, 79–86. doi: 10.1016/j.tics.2003.12.008

Marshall, A. C., Gentsch, A., and Schütz-Bosbach, S. (2018). The interaction between interoceptive and action states within a framework of predictive coding. *Front. Psychol.* 9:180. doi: 10.3389/fpsyg.2018.00180

Maruishi, M., Tanaka, Y., Muranaka, H., Tsuji, T., Ozawa, Y., Imaizumi, S., et al. (2004). Brain activation during manipulation of the myoelectric prosthetic hand: a functional magnetic resonance imaging study. *Neuroimage* 21, 1604–1611. doi: 10.1016/j.neuroimage.2003.12.001

Maselli, A., and Slater, M. (2013). The building blocks of the full body ownership illusion. *Front. Hum. Neurosci.* 7:83. doi: 10.3389/fnhum.2013.00083

Merleau-Ponty, M. (1945). *Phénoménologie de la Perception*. Paris: Gallimard.

Merleau-Ponty, M. (1954). "Eye and mind," in *The Primacy of Perception.* trans. C. Dallery, ed. J. M. Edie (Evanston, IL: Northwestern University Press), 159–190.

Merleau-Ponty, M. (1964). *Le Visible et L'Invisible.* Paris: Tel Gallimard.

Miall, R. C., and Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Netw.* 9, 1265–1279. doi: 10.1016/S0893-6080(96)00035-4

Mocke, V., Weller, L., Frings, C., Rothermund, K., and Kunde, W. (2020). Task relevance determines binding of effect features in action planning. *Attent. Percept. Psychophys.* 82, 3811–3831. doi: 10.3758/s13414-020-02123-x

Murray, C. D. (2004). An interpretative phenomenological analysis of the embodiment of artificial limbs. *Disabil. Rehabil.* 26, 963–973. doi: 10.1080/09638280410001696764

Müsseler, J., and Skottke, E. M. (2011). Compatibility relationships with simple lever tools. *Hum. Fact.* 53, 383–390.

Müsseler, J., and Sutter, C. (2009). Perceiving one's own movements when using a tool. *Conscious. Cogn.* 18, 359–365. doi: 10.1016/j.concog.2009.02.004

Müsseler, J., Kunde, W., Gausepohl, D., and Heuer, H. (2008). Does a tool eliminate spatial compatibility effects? *Eur. J. Cogn. Psychol.* 20, 211–231. doi: 10.1080/09541440701275815

Nabeshima, C., Lungarella, M., and Kuniyoshi, Y. (2005). "Timing-based model of body schema adaptation and its role in perception and tool use: a robot case study," in *Proceedings of the 4th International Conference on Development and Learning, 2005*. (Piscataway, NJ: IEEE), 7–12. doi: 10.1109/DEVLRN.2005.1490935

Nava, E., Gamberini, C., Berardis, A., and Bolognini, N. (2018). Action shapes the sense of body ownership across human development. *Front. Psychol.* 9:2507. doi: 10.3389/fpsyg.2018.02507

O'Regan, J. K. (2011). *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*. New York, NY: Oxford University Press.

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

Page, D. M., George, J. A., Kluger, D. T., Duncan, C., Wendelken, S., Davis, T., et al. (2018). Motor control and sensory feedback enhance prosthesis embodiment and reduce phantom pain after long-term hand amputation. *Front. Hum. Neurosci.* 12:352. doi: 10.3389/fnhum.2018.00352

Pfister, R. (2019). Effect-based action control with body-related effects: implications for empirical approaches to ideomotor action control. *Psychol. Rev.* 126, 153–161. doi: 10.1037/rev0000140

Pritchard, S. C., Zopf, R., Polito, V., Kaplan, D. M., and Williams, M. A. (2016). Non-hierarchical influence of visual form, touch, and position cues on embodiment, agency, and presence in virtual reality. *Front. Psychol.* 7:1649. doi: 10.3389/fpsyg.2016.01649

Ramachandran, V. S. (1998). Consciousness and body image: lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 353, 1851–1859. doi: 10.1098/rstb.1998.0337

Reiss, D., and Marino, L. (2001). Mirror self-recognition in the bottlenose dolphin: a case of cognitive convergence. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5937–5942. doi: 10.1073/pnas.101086398

Renault, A. G., Auvray, M., Parseihian, G., Miall, R. C., Cole, J., and Sarlegna, F. R. (2018). Does proprioception influence human spatial cognition? A study on individuals with massive deafferentation. *Front. Psychol.* 9:1322. doi: 10.3389/fpsyg.2018.01322

Rizzolatti, G. (2005). The mirror neuron system and its function in humans. *Anat. Embryol.* 210, 419–421. doi: 10.1007/s00429-005-0039-z

Rohde, M., Di Luca, M., and Ernst, M. O. (2011). The rubber hand illusion: feeling of ownership and proprioceptive drift do not go hand in hand. *PLoS One* 6:e21659. doi: 10.1371/journal.pone.0021659

Roncone, A., Hoffmann, M., Pattacini, U., and Metta, G. (2014). "Automatic kinematic chain calibration using artificial skin: self-touch in the icub humanoid robot," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*. (Piscataway, NJ: IEEE), 2305–2312. doi: 10.1109/ICRA.2014.6907178

Samad, M., Chung, A. J., and Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLoS One* 10:e0117178. doi: 10.1371/journal.pone.0117178

Schütz-Bosbach, S., and Prinz, W. (2007). Perceptual resonance: action-induced modulation of perception. *Trends Cogn. Sci.* 11, 349–355. doi: 10.1016/j.tics.2007.06.005

Shin, Y. K., Proctor, R. W., and Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychol. Bull.* 136, 943–974. doi: 10.1037/a0020541

Singer, T., and Lamm, C. (2009). The social neuroscience of empathy. *Ann. N. Y. Acad. Sci.* 1156, 81–96.

Slater, M., Spanlang, B., Sanchez-Vives, M. V., and Blanke, O. (2010). First person experience of body transfer in virtual reality. *PLoS One* 5:e10564. doi: 10.1371/journal.pone.0010564

Slaughter, V. (2021). Do newborns have the ability to imitate? *Trends Cogn. Sci.* 25, 377–387. doi: 10.1016/j.tics.2021.02.006

Sterling, P. (2004). "Principles of allostasis: optimal design, predictive regulation, pathophysiology, and rational therapeutics," in *Allostasis, Homeostasis, and the Costs of Physiological Adaptation*, Vol. 17, ed. J. Schulkin (Cambridge, MA: MIT Press), 17–64.

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol. Behav.* 106, 5–15. doi: 10.1016/j.physbeh.2011.06.004

Sülzenbrück, S., and Heuer, H. (2009). Functional independence of explicit and implicit motor adjustments. *Conscious. Cogn.* 18, 145–159. doi: 10.1016/j.concog.2008.12.001

Tajadura-Jiménez, A., Longo, M. R., Coleman, R., and Tsakiris, M. (2012). The person in the mirror: using the enfacement illusion to investigate the experiential structure of self-identification. *Conscious. Cogn.* 21, 1725–1738. doi: 10.1016/j.concog.2012.10.004

Tsakiris, M. (2010). My body in the brain: a neurocognitive model of body-ownership. *Neuropsychologia* 48, 703–712. doi: 10.1016/j.neuropsychologia.2009.09.034

Tsakiris, M. (2017). The multisensory basis of the self: from body to identity to others. *Q. J. Exp. Psychol.* 70, 597–609. doi: 10.1080/17470218.2016.1181768

Verschoor, S. A., and Hommel, B. (2017). Self-by-doing: the role of action for self-acquisition. *Soc. Cogn.* 35, 127–145. doi: 10.1521/soco.2017.35.2.127

Waszak, F., Cardoso-Leite, P., and Hughes, G. (2012). Action effect anticipation: neurophysiological basis and functional consequences. *Neurosci. Biobehav. Rev.* 36, 943–959. doi: 10.1016/j.neubiorev.2011.11.004

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends Cogn. Sci.* 1, 209–216. doi: 10.1016/S1364-6613(97)01070-X

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Curr. Biol.* 11, R729–R732. doi: 10.1016/S0960-9822(01)00432-8

Zaadnoordijk, L., Besold, T. R., and Hunnius, S. (2019). A match does not make a sense: on the sufficiency of the comparator model for explaining the sense of agency. *Neurosci. Conscious.* 2019:niz006. doi: 10.1093/nc/niz006

Zahavi, D. (2001). Beyond empathy. Phenomenological approaches to intersubjectivity. *J. Conscious. Stud.* 8, 151–167.

Zenha, R., Vicente, P., Jamone, L., and Bernardino, A. (2018). “Incremental adaptation of a robot body schema based on touch events,” in *Proceedings of the 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. (Piscataway, NJ: IEEE), 119–124.

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

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

frontiers in Psychology

PERSPECTIVE
published: 26 November 2021
doi: 10.3389/fpsyg.2021.716982

Building and Understanding the Minimal Self

*Valentin Forch and Fred H. Hamker**

Department of Computer Science, Chemnitz University of Technology, Chemnitz, Germany

Within the methodologically diverse interdisciplinary research on the minimal self, we identify two movements with seemingly disparate research agendas – cognitive science and cognitive (developmental) robotics. Cognitive science, on the one hand, devises rather abstract models which can predict and explain human experimental data related to the minimal self. Incorporating the established models of cognitive science and ideas from artificial intelligence, cognitive robotics, on the other hand, aims to build embodied learning machines capable of developing a self "from scratch" similar to human infants. The epistemic promise of the latter approach is that, at some point, robotic models can serve as a testbed for directly investigating the mechanisms that lead to the emergence of the minimal self. While both approaches can be productive for creating causal mechanistic models of the minimal self, we argue that building *a* minimal self is different from understanding *the human* minimal self. Thus, one should be cautious when drawing conclusions about the human minimal self based on robotic model implementations and vice versa. We further point out that incorporating constraints arising from different levels of analysis will be crucial for creating models that can predict, generate, and causally explain behavior in the real world.

Keywords: minimal self, mechanistic models, cognitive robotics, sense of agency, sense of ownership

OPEN ACCESS

Edited by:
Stephan Alexander Verschoor,
Leiden University, Netherlands

Reviewed by:
Anna Belardinelli,
Honda Research Institute Europe GmbH, Germany
Dennis Küster,
University of Bremen, Germany

****Correspondence:***
Fred H. Hamker
fred.hamker@informatik.tu-chemnitz.de

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *29 May 2021*
Accepted: *26 October 2021*
Published: *26 November 2021*

Citation:
Forch V and Hamker FH (2021) Building and Understanding the Minimal Self. Front. Psychol. 12:716982. doi: 10.3389/fpsyg.2021.716982

INTRODUCTION

The minimal self describes the immediate, pre-reflective experience of selfhood derived from sensory information (Gallagher, 2000; Blanke and Metzinger, 2009). Conceptually, it has been subdivided into the sense of agency (SoA, "I produced an outcome with my voluntary action.") and the sense of ownership (SoO, "This body part/mental state belongs to me". Haggard, 2017; Braun et al., 2018). In the wake of experimental paradigms that added implicit measures to the verbally reported experience of SoA (Haggard et al., 2002) and SoO (Botvinick and Cohen, 1998), both concepts have received considerable attention in the behavioral, cognitive, and neurosciences (David et al., 2008; Blanke et al., 2015; Haggard, 2017; Noel et al., 2018). Currently, the field offers a wealth of empirical findings on the antecedents of and relationships among the implicit and explicit behavioral measures of minimal selfhood as well as related neurophysiological measures (see Blanke et al., 2015; Braun et al., 2018; Noel et al., 2018 for reviews).

These advances in the human domain have been paralleled by a growing interest in the different aspects of the minimal self among roboticists and AI researchers who reason that equipping machines with a self-representation similar to humans will ultimately increase their performance and robustness in real-world settings (e.g., Hoffmann et al., 2010; Legaspi et al., 2019;

Hafner et al., 2020). Collaborative efforts of robotics and psychology have been spearheaded by cognitive robotics and further advanced by developmental robotics, which strives for the implementation of a quasi-human developmental scheme for robots (Asada et al., 2009). More specifically, an agentive model embodied by a robot undergoing a developmental phase like human infants could enable direct investigations into the mechanisms that lead to the emergence of a minimal self (Hafner et al., 2020) and thus could be used to test different theories regarding the minimal self.

Current theoretical accounts on the minimal self may be broadly categorized into (a) informal models, including box-and-arrow models and verbal formulations of laws and constraints for the emergence of SoO and SoA (e.g., Synofzik et al., 2008; Tsakiris, 2010; Blanke et al., 2015; Haggard, 2017), (b) Bayesian accounts, according to which the perception of SoO and SoA is governed by statistically optimal information integration, as a main function of the brain is to optimally estimate the state of the world (e.g., Samad et al., 2015; Legaspi and Toyoizumi, 2019), and (c) accounts based on the free energy principle (FEP), which also lends itself to the interpretation of the self as the result of a continuous process of optimizing one's world model (e.g., Limanowski and Blankenburg, 2013; Apps and Tsakiris, 2014; Seth and Friston, 2016).

Much of this theorizing regarding the minimal self is non-mechanistic in the sense that it either focuses on the computational level of cognition (Marr, 1982), which is about describing goals rather than the underlying mechanisms, or does not specify how relevant brain functions are carried out by specific parts of the brain. In more statistical terms, this could be expressed as defining the objective function that needs to be optimized by an agent without specifying the algorithms *the agent* employs to do the optimization. However, if one is interested in building mechanistic models – ones that can causally explain psychological phenomena – it is crucial to account for the algorithmic/representational and implementational levels (Marr, 1982), which describe how and by which parts the goals specified on the computational level are achieved (Piccinini and Craver, 2011; Love, 2015; Kriegeskorte and Douglas, 2018).[1]

The problem of neglecting mechanistic details becomes acute when the use of robotic platforms necessitates model implementation. If a model is underconstrained on the representational and implementational level, researchers will be forced to choose between many algorithms which can achieve the specified computational goal(s; cf. Anderson, 1978). In turn, this is likely to produce a significant deviation of the model from human behavior as not all algorithms for achieving a given computational goal perform equally under non-optimal conditions (e.g., time pressure, insufficient memory capacity, and internal noise) which are characteristic for the real-world settings humans operate in Wang (2019). Moreover, without specifying further constraints, human information integration appears to be non-optimal for many tasks (Rahnev and Denison, 2018; Lieder and Griffiths, 2020). The question of how to reconcile these idiosyncrasies with theories of optimal information integration has sparked an ongoing debate (also see Bowers and Davis, 2012; Griffiths et al., 2012; Love, 2015). In a similar vein, one should consider the context and complexity of the behavior to be modeled (Craver, 2006; Krakauer et al., 2017) – superficial phenomenal descriptions will likely lead to over-simplistic models.

In sum, whatever aspects of the minimal self (or any target system), a model can represent should depend on three factors: (a) the model's objective function or goal (e.g., optimal prediction of the environment and solving a set of tasks), (b) the algorithmic implementation it employs for achieving its goals, and (c) the conditions under which it operates or inputs it receives. We assume that only if all three factors align, the model can serve as a mechanistic explanation. Conversely, if mechanistic details are not specified and phenomenal similarities between humans and robots are superficial, drawing conclusions from model implementations to humans (and vice versa) would be ill-advised.

Thus, the present contribution aims at highlighting the need for deeper integration of insights from the behavioral, cognitive, and neurosciences if one's goal is a better understanding of the human minimal self. Of course, the interactive approach of robotics and ideas from artificial intelligence benefit cognitive neuroscience (Marblestone et al., 2016; Hoffmann and Pfeifer, 2018). We contend, however, that only models of the human minimal self which are phenomenologically rich and specify mechanistic details can be meaningfully tested through robotic model implementations. In the remainder, we will go into more detail regarding (a) the role of causal mechanistic models in cognitive neuroscience, (b) the mechanistic depth of different models of aspects of the minimal self, and (c) the current state of cognitive and developmental robotics implementations of such models.

CAUSAL MECHANISTIC MODELS IN COGNITIVE NEUROSCIENCE

Understanding a phenomenon requires being able to explain how said phenomenon comes about (or fails to do so) under certain circumstances. Such causal explanations need to specify the mechanism producing said phenomenon (Craver, 2006). A mechanism is defined as being composed of parts whose organized activity produces a phenomenon from certain starting conditions (Machamer et al., 2000; Craver, 2006). Crucially, there needs to be a clear relation between parts and processes (Hommel, 2020) and the assumed parts of the mechanism need to be measurable and open to intervention to make the causal model testable (Craver, 2006).

The notion of causal mechanistic models does not imply reductionism (Nicholson, 2012), that is, that human behavior can be explained satisfactorily in the language of neuroscience, molecular biology, or particle physics alone. Rather, it is open to multilevel explanations (Kaplan and Craver, 2011). Crucially, this also requires a thorough description of the phenomenon

[1] When talking about the implementational level, we do not exclusively refer to singular neurons or synapses. Groups of neurons or brain areas may also be related to a function. To be verifiable mechanistic parts, the states of such a physical system still need to be measurable and clearly attributable to the implementation of a concrete algorithm.

to be explained and a distinction between standard and non-standard (e.g., lab) conditions (Craver, 2006). If the conditions under which a phenomenon is observed and described are non-representative of the real world, a model trying to explain it will likely not generalize well to real-world scenarios. Models in (computational) neuroscience have been criticized for being too reductionist, focusing on biological mechanisms that cannot be related to meaningful behavior (Krakauer et al., 2017).

Descriptive models, on the other hand, act as a compact summary of a phenomenon (Kaplan and Craver, 2011). They enable predictions about the phenomenon, without specifying the underlying mechanism. This type of model is widespread in psychology and cognitive neuroscience (Kaplan and Craver, 2011; Hommel, 2020; Litwin and Miłkowski, 2020) and can be derived from general assumptions about brain function (e.g., "the brain optimizes an internal world model") or empirical observations (e.g., the rubber hand illusion, brain imaging data). A descriptive model can still serve as a starting point for building a causal model if it is possible to relate parts of the model to parts of a causal mechanism (Kaplan and Craver, 2011; Piccinini and Craver, 2011). Moreover, in the face of physiological and behavioral complexity, the notion of a truly mechanistic model appears somewhat idealized and may be only approached gradually, making descriptive models a reasonable starting point.

MECHANISTIC DEPTH OF MODELS OF THE MINIMAL SELF

Starting with informal descriptive models of the minimal self, we will consider the work by Tsakiris (2010) (see also Wegner and Wheatley, 1999; Frith et al., 2000; Synofzik et al., 2008; Chambon et al., 2014; Blanke et al., 2015). This model is concerned with explaining the SoO over body parts or objects. It proposes a tiered comparison between the features of candidate objects for experiencing ownership and the current state of an internal body model (i.e., comparison of visual appearance, posture, and sensory stimulation – in this order). Tsakiris (2010) also points toward evidence of certain brain areas being responsible for this comparison. While the model provides an algorithm in the sense that it specifies the order in which certain information is compared, it includes no constraints on the algorithms for making the comparisons or how they could be implemented by the brain. It also does not specify how the internal model of the body is represented.

Although the model makes testable predictions, it is clearly not mechanistic to the degree that it would permit a straightforward robotic implementation without additional assumptions. The same holds for other informal models which specify what kind of information is processed, but which do not provide the actual metric used for making comparisons or the processes underlying the formation of representations. **Figure 1** tries to make a graphical comparison between the human self-representation and models of the human self. Informal models typically account for relatively broad phenomena like the SoO. Thus, they cover a large part of the human "self-space" (observable self-related behaviors and self-related information relevant for constructing the internal self-representation). However, as they are only loosely constrained by theoretical assumptions and do not make quantifiable predictions, these models would likely conform with behavior that is outside the human repertoire.

Bayesian models (e.g., Samad et al., 2015; Legaspi and Toyoizumi, 2019) frame the perception of SoA and SoO as the posterior probability for perceiving objects or actions as belonging to or being caused by oneself given sensory input and prior beliefs. These models can be very useful for untangling what information is relevant for a certain task or percept (e.g., Legaspi and Toyoizumi, 2019) but usually make no commitments to the algorithms employed by the brain (Griffiths et al., 2012; Love, 2015). Neurocomputational models for approximating Bayesian inference (e.g., Pouget et al., 2000) try to build a bridge between computational goals and concrete implementations (cf. Love, 2015) and have been shown to fit the response characteristics of biological neurons (Avillac et al., 2005).

While neurocomputational models for multisensory integration – which is thought to be central for the SoO – are abundant (see Ursino et al., 2014; Blanke et al., 2015 for reviews), there are still explanatory gaps: (a) many of these models feature no learning mechanism (e.g., Deneve et al., 2001) or use learning techniques that cannot be brought into correspondence with parts and processes of the brain (i.e., the use of machine learning techniques, Makin et al., 2013), (b) many models are based on physiological data from midbrain structures (e.g., Cuppini et al., 2012; Oess et al., 2020), where the empirical link between these structures and the perception of SoO is not clear, and (c) the neurophysiological constraints incorporated into these models so far have not been demonstrated to give rise to more specific predictions on the behavioral level.

The latter point is important because traditional Bayesian models and thus their neurocomputational counterparts often only apply to human behavior in idealized situations (Love, 2015; Rahnev and Denison, 2018). Research from other domains, however, has shown that taking additional constraints on the representational level (e.g., efficient coding; Wei and Stocker, 2015) or implementational level (e.g., internal noise; Tsetsos et al., 2016) into account can greatly benefit modeling "non-optimal" human behavior in real-world settings (also see Lieder and Griffiths, 2020 for a review). These examples show that by refining computational models with more low-level constraints instead of simply translating them into a neurocomputational framework, it is possible to move closer to the human style of information processing – also an exciting opportunity for research on the self.

The FEP builds on the notion that human brains, like all living systems, can be thought of as "trying" to minimize their surprisal through representing an optimal world model and acting on it (Friston, 2010). At its core, the FEP is closely related to Bayesianism (Aitchison and Lengyel, 2017) but incorporates a (variable) host of additional assumptions (Gershman, 2019; Bruineberg et al., 2020), the most important arguably

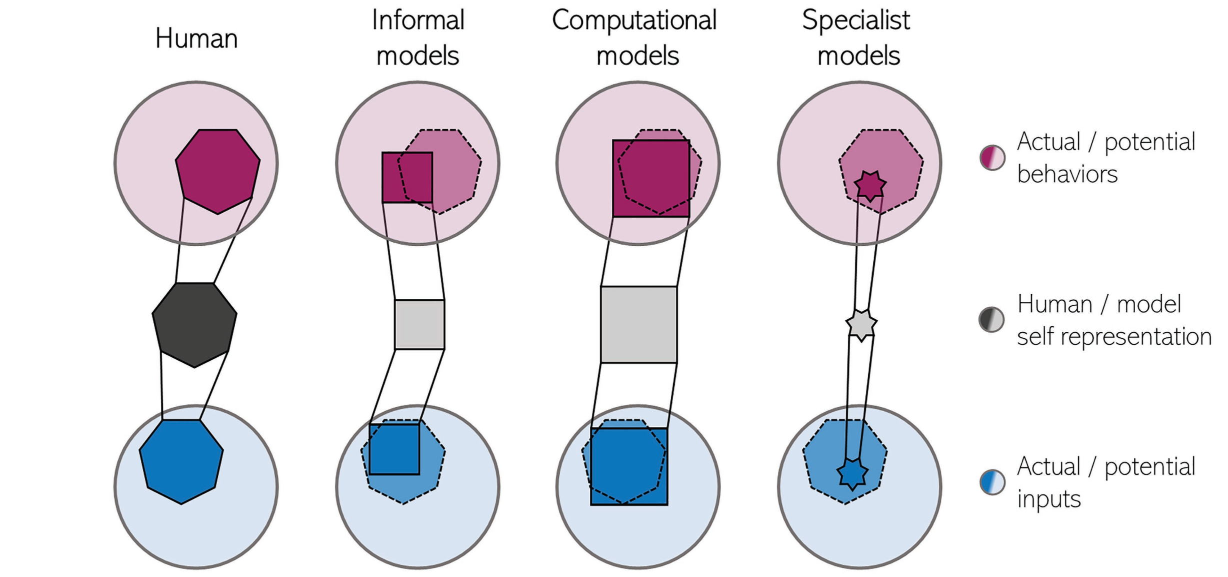

Human
Informal models
Computational models
Specialist models
Actual / potential behaviors
Human / model self representation
Actual / potential inputs

FIGURE 1 | Schematic of the representational power of models for the human minimal self. The human minimal self representation (dark grey, left) is based on a relevant subset of all self-related information (blue) while also underlying a subset of all possible self-related behaviors (purple, e.g., experiencing a piece of furniture as belonging to oneself). Currently, models of the self (right) are too narrow in the sense that they consider only a subset of potential inputs and in practice can generate only a small subset of human behaviors and/or too general in the sense that they make too unspecific predictions regarding self-related phenomena, violating the bounds of the human self (dotted line). Note, that the relative number of constraints underlying each self-representation is reflected by the number of sides of the respective shapes.

being the explicit representation of prediction errors at all stages of perception and action, termed predictive coding (PC, Rao and Ballard, 1999; see Aitchison and Lengyel, 2017 for PC schemes in other contexts). According to PC, predictions descend the cortical hierarchy where they suppress incoming bottom-up signals leading to the representation of prediction errors. These prediction errors, in turn, are propagated up the hierarchy to inform the update of higher-level representations. Ultimately, this leads to a dynamic equilibrium where prediction errors are minimized (Friston, 2010).

The FEP and PC have been rapidly adopted in the domains of interoception and the (minimal) self (e.g., Limanowski and Blankenburg, 2013; Apps and Tsakiris, 2014; Barrett and Simmons, 2015; Seth and Friston, 2016). Building on PC, Apps and Tsakiris (2014), for instance, explain illusions of ownership over extracorporeal objects like the rubber hand illusion as a process where prediction errors caused by incongruent sensory information are "explained away" by updating one's high-level representations in such a way that best predicts said sensory information. However, the authors do not specify how the prediction errors are computed or how they are transformed into beliefs.

This gap may be closed by neurocomputational models of PC (Bastos et al., 2012). However, as neurophysiological evidence for PC is inconclusive (Seth and Friston, 2016; Aitchison and Lengyel, 2017), this vein of research requires further investigation (Keller and Mrsic-Flogel, 2018). Additionally, the same reservation as for Bayesian models applies – in our view, showing that an optimization scheme can be implemented through neural computation, while being necessary for a possible mechanistic explanation, is not sufficient as long as the more specific model does not capture relevant deviations from behavior predicted by computational constraints alone.

One such deviation yet unexplained by computational models may be the apparent dissociation of explicit and implicit measures of SoO in the rubber hand illusion under certain conditions (Holle et al., 2011; Rohde et al., 2011; Gallagher et al., 2021), which has been explained under the same framework of information integration (Apps and Tsakiris, 2014). Another example is the effect of action selection fluency on SoA (Chambon et al., 2014) which shows that the SoA can be diminished solely by hindering fluent action selection. This effect is independent of the predictability of the action outcome – the core tenet of comparator models of SoA (Frith et al., 2000) which strongly align with PC (cf. Aitchison and Lengyel, 2017). Coming back to **Figure 1**, we would then argue that, albeit being very broad in scope, computational models of the minimal self are only a first approximation of the information processing underlying the minimal self. Refining these models with new constraints will necessitate synergistic modeling and empirical work – behavioral scientists will have to further explore the limits of the malleability of the human minimal self and the relative importance of different kinds of information used for constructing it, thereby informing theorists who, in turn, should create models that make new, empirically testable predictions, thus entering an experiment-model development-prediction cycle of research. One concrete future direction might be considering multiple computational constraints which could even play different roles during development (cf. Marblestone et al., 2016). Besides prediction error reduction this could be, for instance, novelty, reward maximization, or computational efficiency.

MINIMAL SELF-MODELS IN COGNITIVE AND DEVELOPMENTAL ROBOTICS

Applying a theory or model in a complex environment either through simulation or the use of physical robots may speed up research efforts significantly by reducing the need for time-consuming human experiments and increasing the control and transparency of the subject. Unfortunately, reviewing robotic models related to the minimal self would be beyond the scope of this contribution (see Nguyen et al., 2021 for an excellent review). Instead, we want to point out two tendencies that may impair the epistemic power of robotic model implementations.

Compared to traditional cognitive and neuroscience models, robotic implementations have the advantage of receiving rather realistic input as robots can directly interact with the real world and register the consequences of their actions (Hoffmann and Pfeifer, 2018). Moreover, the use of embodied agents allows testing the impact of physiological features (i.e., body morphology) on learned representations. This increased fidelity of model inputs, however, makes implementations much more demanding. Thus, it is not surprising that robotic model implementations often rely on more scalable machine learning techniques instead of neurocomputational models (cf. Nguyen et al., 2021). This has the benefit of introducing powerful ideas like curiosity-driven learning (Oudeyer et al., 2007), but also contains the risk of deviating on the algorithmic level by choosing an algorithm that elegantly solves a given task while neglecting biological constraints. We assume this concern will bear greater importance when task complexity increases and experimental settings move closer toward the real world.

Second, as Krichmar (2012) noted, cognitive robotics models, in general, tend to be built to perform very specific tasks. This diminishes the ecological benefit of real-world inputs because it greatly reduces the possible robot-world interactions. Moreover, the use of narrow tasks holds the risk of over-engineering the model to the task (as, e.g., Hoffmann et al. (2021) note for robotic models of minimal self-awareness). Such specialist models will hardly generalize in novel situations. Covering the whole self-space (**Figure 1**) would then require a multitude of such models that need to be integrated somehow, which would be a daunting task (Clune, 2020). Moreover, testing a robotic implementation under quasi-lab conditions only for the behaviors which have been used to build and train the underlying model cannot be regarded as a critical test of a theory.

One promising approach, therefore, appears to be letting robots solve general tasks that necessitate real-world interactions without explicitly engineering the model to perform a specific behavior, like say, attenuating self-caused sensory input – which has been related to SoA (Schillaci et al., 2016; but see Kaiser and Schütz-Bosbach, 2018). In such a scenario, the robot should show some behavior because it is (a) possible and (b) beneficial for task success. One could then proceed by probing the conditions under which this behavior develops or is enacted. By comparing the model to human behavior under diverse conditions, one could simultaneously test the assumed mechanism and deepen the phenomenological description of the human repertoire. This method could even be generalized to the point where the agent is not designed by the researcher but by an (evolutionary) algorithm guided by task success and prior constraints (cf. Albantakis et al., 2014). However, such an approach might require going to the edge of what is currently computationally possible (cf. Clune, 2020).

DISCUSSION: WHY MECHANISTIC MODELS?

So far, we have established that there is no complete mechanistic explanation of the minimal self yet – but why should mechanistic models be beneficial for further research on the minimal self? We see several benefits in striving for integrating evidence from different levels of description and thereby creating more mechanistic models of the minimal self: (a) It safeguards against overfitting to specific pieces of evidence, assumptions, or tasks, (b) it increases model comparability and the probability of model generalization, and (c) especially in clinical contexts, a causal understanding may help to find effective interventions for (self-)disorders and interfaces with other theories (e.g., Schroll and Hamker, 2016; Neumann et al., 2018). For brevity, we will only touch upon the first two points.

Anchoring a model in a narrow set of observations, assumptions, or tasks bears the risk of selectively including evidence that fits the model and tailoring the model to these data points (cf. Love, 2015). Because mechanistic models demand a multilevel view on a phenomenon, their implementation should counteract this risk. They should also increase model comparability as there can be no meaningful comparison of two models that make predictions for distinct variables or solve different tasks (Love, 2021). As the minimal self and its subcomponents are relevant in many contexts, their corresponding mechanistic models should also not be bound to a narrow task.

Furthermore, explicitly distinguishing between mechanistic and non-mechanistic models also helps when thinking about robots as models for the human minimal self. If we understand the self as a representation of contextually and ethologically relevant features of one's physical body and intentional actions which is learned and continuously updated by the nervous system, we may ascribe a minimal (pre-reflective) self to very primitive creatures like ants. Ants have been shown to perform approximately optimal cue integration of vision and proprioception (Wystrach et al., 2015),[2] act intentionally (Hunt et al., 2016), and learn (Dupuy et al., 2006). Admittedly being an exaggeration, this example should make clear that if we exclude higher-order cognition (as it is not pre-reflective), ignore individual representational capacities, behavioral complexity, and other conditions constraining sensory content,

[2] The study of Wystrach et al. (2015) also provides an example of the importance of implementation constraints affecting behavior. Ants show "suboptimal" cue integration under some circumstances which could be explained by a memory restriction in their information processing.

we run the risk of ascribing some phenomenology to systems vastly different from us.

Certainly, there is much potential in using embodied machines to advance investigations into the human minimal self. However, we would caution against thinking of both as being representative for one another as long as there is no agreement between all levels of description relevant for cognition and behavior. This should not imply that robot "brains" or other models need to be neuromorphic, but as the human brain is a product of the chaotic process of evolution, and given that there is no unique implementation of purely computational theories due to the complexity and dynamics of real-world settings (Whiteley and Sahani, 2012; Gershman, 2019), it appears unlikely that an algorithm that is only constrained by a single computational goal could fully capture human behavior and experience (cf. Marblestone et al., 2016; Kriegeskorte and Douglas, 2018; Lieder and Griffiths, 2020). In conclusion, incorporating constraints arising from different levels of analysis will be crucial for creating models able to predict, generate, and mechanistically explain behavior related to the minimal self in the real world.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

VF and FH jointly developed the general idea discussed in the manuscript. VF wrote the manuscript and produced **Figure 1**. FH reviewed the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the DFG priority program "The Active Self" HA2630/12-1.

REFERENCES

Aitchison, L., and Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Curr. Opin. Neurobiol.* 46, 219–227. doi: 10.1016/j.conb.2017.08.010

Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput. Biol.* 10:e1003966. doi: 0.1371/journal.pcbi.1003966

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychol. Rev.* 85, 249–277. doi: 10.1037/0033-295X.85.4.249

Apps, M. A., and Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neurosci. Biobehav. Rev.* 41, 85–97. doi: 10.1016/j.neubiorev.2013.01.029

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

Avillac, M., Deneve, S., Olivier, E., Pouget, A., and Duhamel, J. R. (2005). Reference frames for representing visual and tactile locations in parietal cortex. *Nat. Neurosci.* 8, 941–949. doi: 10.1038/nn1480

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429. doi: 10.1038/nrn3950

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Blanke, O., and Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends Cogn. Sci.* 13, 7–13. doi: 10.1016/j.tics.2008.10.003

Blanke, O., Slater, M., and Serino, A. (2015). Behavioral, neural, and computational principles of bodily self-consciousness. *Neuron* 88, 145–166. doi: 10.1016/j.neuron.2015.09.029

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature* 391:756. doi: 10.1038/35784

Bowers, J. S., and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* 138:389. doi: 10.1037/a0026450

Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H., et al. (2018). The senses of agency and ownership: a review. *Front. Psychol.* 9:535. doi: 10.3389/fpsyg.2018.00535

Bruineberg, J., Dolega, K., Dewhurst, J., and Baltieri, M. (2020). The emperor's new Markov blankets [Preprint]. Available at: http://philsci-archive.pitt.edu/id/eprint/18467 (Accessed March 20, 2021).

Chambon, V., Sidarus, N., and Haggard, P. (2014). From action intentions to action effects: how does the sense of agency come about? *Front. Hum. Neurosci.* 8:320. doi: 10.3389/fnhum.2014.00320

Clune, J. (2020). AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. arXiv [Preprint]. Available at: https://arxiv.org/abs/1905.10985 (Accessed April 14, 2021).

Craver, C. F. (2006). When mechanistic models explain. *Synthese* 153, 355–376. doi: 10.1007/s11229-006-9097-x

Cuppini, C., Magosso, E., Rowland, B., Stein, B., and Ursino, M. (2012). Hebbian mechanisms help explain development of multisensory integration in the superior colliculus: a neural network model. *Biol. Cybern.* 106, 691–713. doi: 10.1007/s00422-012-0511-9

David, N., Newen, A., and Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Conscious. Cogn.* 17, 523–534. doi: 10.1016/j.concog.2008.03.004

Deneve, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat. Neurosci.* 4, 826–831. doi: 10.1038/90541

Dupuy, F., Sandoz, J. C., Giurfa, M., and Josens, R. (2006). Individual olfactory learning in Camponotus ants. *Anim. Behav.* 72, 1081–1091. doi: 10.1016/j.anbehav.2006.03.011

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

Frith, C. D., Blakemore, S.-J., and Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. *Brain Res. Rev.* 31, 357–363. doi: 10.1016/S0165-0173(99)00052-1

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Gallagher, M., Colzi, C., and Sedda, A. (2021). Dissociation of proprioceptive drift and feelings of ownership in the somatic rubber hand illusion. *Acta Psychol.* 212:103192. doi: 10.1016/j.actpsy.2020.103192

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? arXiv [Preprint]. Available at: https://arxiv.org/abs/1901.07945 (Accessed March 15, 2021).

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). *Psychol. Bull.* 138, 415–422. doi: 10.1037/a0026884

Hafner, V. V., Loviken, P., Villalpando, A. P., and Schillaci, G. (2020). Prerequisites for an artificial self. *Front. Neurorobot.* 14:5. doi: 10.3389/fnbot.2020.00005

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 196–207. doi: 10.1038/nrn.2017.14

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nat. Neurosci.* 5, 382–385. doi: 10.1038/nn827

Hoffmann, M., Marques, H., Arieta, A., Sumioka, H., Lungarella, M., and Pfeifer, R. (2010). Body schema in robotics: a review. *IEEE Trans. Auton. Ment. Dev.* 2, 304–324. doi: 10.1109/TAMD.2010.2086454

Hoffmann, M., and Pfeifer, R. (2018). "Robots as powerful allies for the study of embodied cognition from the bottom up," in *The Oxford Handbook of 4e Cognition*, eds. A. Newen, L. de Bruin and S. Gallagher (New York: Oxford University Press), 841–862.

Hoffmann, M., Wang, S., Outrata, V., Alzueta, E., and Lanillos, P. (2021). Robot in the mirror: toward an embodied computational model of mirror self-recognition. *KI-Künstl. Int.* 35, 37–51. doi: 10.1007/s13218-020-00701-7

Holle, H., McLatchie, N., Maurer, S., and Ward, J. (2011). Proprioceptive drift without illusions of ownership for rotated hands in the "rubber hand illusion" paradigm. *Cogn. Neurosci.* 2, 171–178. doi: 10.1080/17588928.2011.603828

Hommel, B. (2020). Pseudo-mechanistic explanations in psychology and cognitive neuroscience. *Top. Cogn. Sci.* 12, 1294–1305. doi: 10.1111/tops.12448

Hunt, E. R., Baddeley, R. J., Worley, A., Sendova-Franks, A. B., and Franks, N. R. (2016). Ants determine their next move at rest: motor planning and causality in complex systems. *R. Soc. Open Sci.* 3:150534. doi: 10.1098/rsos.150534

Kaiser, J., and Schütz-Bosbach, S. (2018). Sensory attenuation of self-produced signals does not rely on self-specific motor predictions. *Eur. J. Neurosci.* 47, 1303–1310. doi: 10.1111/ejn.13931

Kaplan, D. M., and Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philos. Sci.* 78, 601–627. doi: 10.1086/661755

Keller, G. B., and Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron* 100, 424–435. doi: 10.1016/j.neuron.2018.10.003

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Krichmar, J. L. (2012). Design principles for biologically inspired cognitive robotics. *Biol. Inspired Cogn. Archit.* 1, 73–81. doi: 10.1016/j.bica.2012.04.003

Kriegeskorte, N., and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nat. Neurosci.* 21, 1148–1160. doi: 10.1038/s41593-018-0210-5

Legaspi, R., He, Z., and Toyoizumi, T. (2019). Synthetic agency: sense of agency in artificial intelligence. *Curr. Opin. Behav. Sci.* 29, 84–90. doi: 10.1016/j.cobeha.2019.04.004

Legaspi, R., and Toyoizumi, T. (2019). A Bayesian psychophysics model of sense of agency. *Nat. Commun.* 10, 1–11. doi: 10.1038/s41467-019-12170-0

Lieder, F., and Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43:e1. doi: 10.1017/S0140525X1900061X

Limanowski, J., and Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Front. Hum. Neurosci.* 7:547. doi: 10.3389/fnhum.2013.00547

Litwin, P., and Miłkowski, M. (2020). Unification by fiat: arrested development of predictive processing. *Cogn. Sci.* 44:e12867. doi: 10.1111/cogs.12867

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Top. Cogn. Sci.* 7, 230–242. doi: 10.1111/tops.12131

Love, B. C. (2021). Levels of biological plausibility. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 376:20190632. doi: 10.1098/rstb.2019.0632

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philos. Sci.* 67, 1–25. doi: 10.1086/392759

Makin, J. G., Fellows, M. R., and Sabes, P. N. (2013). Learning multisensory integration and coordinate transformation via density estimation. *PLoS Comput. Biol.* 9:e1003035. doi: 10.1371/journal.pcbi.1003035

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10:94. doi: 10.3389/fncom.2016.00094

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT press.

Neumann, W. J., Schroll, H., de Almeida Marcelino, A. L., Horn, A., Ewert, S., Irmen, F., et al. (2018). Functional segregation of basal ganglia pathways in Parkinson's disease. *Brain* 141, 2655–2669. doi: 10.1093/brain/awy206

Nguyen, P. D., Georgie, Y. K., Kayhan, E., Eppe, M., Hafner, V. V., and Wermter, S. (2021). Sensorimotor representation learning for an "active self" in robots: a model survey. *KI-Künstl. Int.* 35, 9–35. doi: 10.1007/s13218-021-00703-z

Nicholson, D. J. (2012). The concept of mechanism in biology. *Stud. Hist. Phil. Biol. Biomed. Sci.* 43, 152–163. doi: 10.1016/j.shpsc.2011.05.014

Noel, J. P., Blanke, O., and Serino, A. (2018). From multisensory integration in peripersonal space to bodily self-consciousness: from statistical regularities to statistical inference. *Ann. N. Y. Acad. Sci.* 1426, 146–165. doi: 10.1111/nyas.13867

Oess, T., Löhr, M. P., Schmid, D., Ernst, M. O., and Neumann, H. (2020). From near-optimal bayesian integration to neuromorphic hardware: a neural network model of multisensory integration. *Front. Neurorobot.* 14:29. doi: 10.3389/fnbot.2020.00029

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

Piccinini, G., and Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183, 283–311. doi: 10.1007/s11229-011-9898-4

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132. doi: 10.1038/35039062

Rahnev, D., and Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behav. Brain Sci.* 41:e223. doi: 10.1017/S0140525X18000936

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rohde, M., Di Luca, M., and Ernst, M. O. (2011). The rubber hand illusion: feeling of ownership and proprioceptive drift do not go hand in hand. *PLoS One* 6:e21659. doi: 10.1371/journal.pone.0021659

Samad, M., Chung, A. J., and Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLoS One* 10:e0117178. doi: 10.1371/journal.pone.0117178

Schillaci, G., Ritter, C. N., Hafner, V. V., and Lara, B. (2016). "Body representations for robot ego-noise modelling and prediction. Towards the development of a sense of agency in artificial agents." in *International Conference on the Simulation and Synthesis of Living Systems (ALife XV) (Cancún)*; July 4–6, 2016.

Schroll, H., and Hamker, F. H. (2016). Basal ganglia dysfunctions in movement disorders: what can be learned from computational simulations. *Mov. Disord.* 31, 1591–1601. doi: 10.1002/mds.26719

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

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* 17, 219–239. doi: 10.1016/j.concog.2007.03.010

Tsakiris, M. (2010). My body in the brain: a neurocognitive model of body-ownership. *Neuropsychiatrie* 48, 703–712. doi: 10.1016/j.neuropsychologia.2009.09.034

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., and Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proc. Natl. Acad. Sci.* 113, 3102–3107. doi: 10.1073/pnas.1519157113

Ursino, M., Cuppini, C., and Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Netw.* 60, 141–165. doi: 10.1016/j.neunet.2014.08.003

Wang, P. (2019). On defining artificial intelligence. *J. Artif. Gen. Int.* 10, 1–37. doi: 10.2478/jagi-2019-0002

Wegner, D. M., and Wheatley, T. (1999). Apparent mental causation: sources of the experience of will. *Am. Psychol.* 54:480. doi: 10.1037/0003-066X.54.7.480

Wei, X. X., and Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* 18:1509. doi: 10.1038/nn.4105

Whiteley, L., and Sahani, M. (2012). Attention in a Bayesian framework. *Front. Hum. Neurosci.* 6:100. doi: 10.3389/fnhum.2012.00100

Wystrach, A., Mangan, M., and Webb, B. (2015). Optimal cue integration in ants. *Proc. R. Soc. B Biol. Sci.* 282:20151484. doi: 10.1098/rspb.2015.1484

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

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

ORIGINAL RESEARCH
published: 29 November 2021
doi: 10.3389/fpsyg.2021.716433

How Morphological Computation Shapes Integrated Information in Embodied Agents

Carlotta Langer [1,2]* *and Nihat Ay* [1,2,3,4]

[1] Hamburg University of Technology, Hamburg, Germany, [2] Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, [3] Santa Fe Institute, Santa Fe, NM, United States, [4] Leipzig University, Leipzig, Germany

The Integrated Information Theory provides a quantitative approach to consciousness and can be applied to neural networks. An embodied agent controlled by such a network influences and is being influenced by its environment. This involves, on the one hand, morphological computation within goal directed action and, on the other hand, integrated information within the controller, the agent's brain. In this article, we combine different methods in order to examine the information flows among and within the body, the brain and the environment of an agent. This allows us to relate various information flows to each other. We test this framework in a simple experimental setup. There, we calculate the optimal policy for goal-directed behavior based on the "planning as inference" method, in which the information-geometric em-algorithm is used to optimize the likelihood of the goal. Morphological computation and integrated information are then calculated with respect to the optimal policies. Comparing the dynamics of these measures under changing morphological circumstances highlights the antagonistic relationship between these two concepts. The more morphological computation is involved, the less information integration within the brain is required. In order to determine the influence of the brain on the behavior of the agent it is necessary to additionally measure the information flow to and from the brain.

Keywords: information theory, information geometry, planning as inference, morphological computation, integrated information, embodied artificial intelligence

OPEN ACCESS

Edited by:
Verena V. Hafner,
Humboldt University of Berlin,
Germany

Reviewed by:
Astrid Zeman,
KU Leuven, Belgium
Dominik M. Endres,
University of Marburg, Germany

****Correspondence:***
Carlotta Langer
carlotta.langer@tuhh.de

Specialty section:
This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

Received: *01 June 2021*
Accepted: *27 October 2021*
Published: *29 November 2021*

Citation:
Langer C and Ay N (2021) How
Morphological Computation Shapes
Integrated Information in Embodied
Agents. Front. Psychol. 12:716433.
doi: 10.3389/fpsyg.2021.716433

1. INTRODUCTION

1.1. Objective

An agent that is faced with a task can solve it using solely its brain, its body's interaction with the world, or a combination of both. This article presents a framework to analyze the importance of these different interactions for an embodied agent and therefore aims at advancing the understanding of how embodiment influences the brain and the behavior of an agent. To illustrate the idea we discuss the following scenario:

Consider a sailor at sea without any navigational equipment. The sailor has to rely on the information given by the sun or the visible stars in order to determine in which direction to steer. The more complex part of the task is solved by the information processed in the brain of the sailor. On the other hand, a bird equipped with magneto-reception, meaning one that is able to use the magnetic field of the earth to perceive its direction, can rely on this sense and does not need to integrate different sources of information. Here, the body of the bird interacts with the environment for the bird to orientate itself. The complexity of the task is met by the morphology of the bird. Taking this example further we consider a modern boat with a highly developed navigation system. The sailor now only needs to know how to interpret the machines and will therefore have less complex calculations to do. The complexity of the task shifts from the brain and background knowledge of the sailor toward the construction of the navigation system, which receives and integrates different information sources for the sailor to use.

Our objective is to analyze these shifts of complexity. We will do that by quantifying the importance of the information flow in an embodied agent performing a task under different morphological circumstances.

The importance of the human body for perception of the environment and ourselves is a core idea of the embodied cognition theory, see for example (Wilson, 2002) or (Gallagher, 2005). In Gallagher (2000) the author develops a definition of a human minimal self in the following way:

> "Phenomenologically, that is, in terms of how one experiences it, a consciousness of oneself as an immediate subject of experience, unextended in time. The minimal self almost certainly depends on brain processes and an ecologically embedded body, but one does not have to know or be aware of this to have an experience that still counts as a self-experience."

Therefore, it is important to understand the influence the ecologically embedded body has on the brain. Hence, here we aim at quantifying both, the interaction of the body with the environment and the information flows inside the body and the brain, respectively, using the same framework and thereby relating them to each other. As a first step in that direction we will analyze simulated artificial agents in a toy example. These agents have a control architecture, the brain of the agent, consisting of a neural network. This will provide the basis for future analysis of more complex agents such as humanoid robots. Ultimately, we hope to gain insights about human agency, and in particular the representation of the self.

The setting of our experiment will be presented in section 2.1. The question we ask is: How is the complexity of solving the task distributed among the different parts of the body, brain and environment?

The main statements that we will support by our experiments are:

1. The more the agent can rely on the interaction of its body with the environment to solve a task, the less integrated information in the brain is required.

This antagonistic relationship between integrated information and morphological computation can be observed even in cases in which the controller has no influence on the behavior of the agent. Hence it is necessary to analyze further information flows in order to fully understand the impact of the controller on the behavior.

2. The importance of integrated information in the controller for the behavior of an embodied agent depends additionally on the information flowing to and from the controller. Therefore, it is not sufficient to only calculate an integrated information measure for understanding its behavioral implications.

In order to test these statements, we need to develop a theoretical background.

1.2. Theoretical Background

We will model the different interactions using the sensori-motor loop, which depicts the connections among the world W, the controller C, the sensors S and actuators A. This will be discussed further in section 2.1.2.

Using the sensori-motor loop we are able to define a set of probability distributions reflecting the structure of the information flow of an agent interacting with the world. Now we need to find the probability distributions that describe a behavior that optimizes the likelihood of success. It would be possible to use a learning or evolutionary algorithm on the agents to find this optimal behavior, but instead we will apply a method called "planning as inference."

Planning as inference is a technique proposed in Attias (2003), in which a goal directed planning task under uncertainty is solved by probabilistic inference tools. This method models the actions an agent can perform as latent variables. These variables are then optimized with respect to a goal variable using the em-algorithm, an information geometric algorithm that is guaranteed to converge, as proven in Amari (1995). This algorithm might result in local minima depending on the input distribution, which allows us to analyze different kinds of agents and strategies that lead to a similar probability of success. This course of action has the advantage that we can directly calculate the optimal policies without having to first train the agents. We will describe this method in the context of our experimental setup in further detail in section 2.2.

Having calculated the distributions that describe the optimal behavior, we apply various information theoretic measures to quantify the strength of the different connections. The measures we are going to discuss are defined by minimizing the KL-divergence between the original distribution and the set of split distributions. The split distributions lack the information flow that we want to measure. Following this concept we are able to quantify the strength of the different information flows, which leads to measures that can be interpreted as integrated information and morphological computation, respectively. We will further define four additional measures that together quantify all the connections among the controller, sensors and actuators. These are defined in section 2.3.

Using information theoretic measures to quantify the information flow in an embodied agent is a natural approach, since we could perceive the different parts of the system as communicating with each other. Surely the world does not actively send information to the controller, but the controller still receives information about the world through the sensors. There have been various studies analyzing acting agents by using information theoretic measures. In Klyubin et al. (2007) maximizing the information flow through the whole system is used as a learning objective. Furthermore, in Touchette and

Lloyd (2004) the authors use the concepts of information and entropy to define conditions under which a system is perfectly controllable or observable. Emphasizing the importance of the sensory input, entropy and mutual information are utilized in Sporns and Pegors (2004) to analyze how an agent actively structures its sensory input. Moreover, the authors of Lungarella et al. (2005) also include the structure of the motor data in their analysis. The last two cited articles additionally discuss two measures regarding the amount of information and the complexity of its integration. These concepts are also important in the context of Integrated Information Theory.

Integrated Information Theory (IIT) proposed by Tononi aims at measuring the amount and quality of consciousness. This theory went through multiple phases of development starting as a measure for brain complexity (Tononi et al., 1994) and then evolved through different iterations (Tononi and Edelman, 1998; Tononi, 2008), toward a broad theory of consciousness (Oizumi et al., 2014). The two key concepts that are present in all versions of IIT are "Information" and "Integration." Information refers to the number of states a system can be in and Integration describes the amount to which the information is integrated among the different parts of it. Measures for integrated information differ depending on the version of the theory they are referring to and on the framework they are defined in. We discussed a branch of these measures building on information geometry in Langer and Ay (2020). In this article we will use the measure that we propose in Langer and Ay (2020) in the case of a known environment, as defined in section 2.3.1. As advocated by the authors of Mediano et al. (2021) we will treat the integrated information measure as a complexity measure and therefore as a way to quantify the relevant information flow in the controller.

Another general feature of all IIT measures so far is that they focus solely on the brain, meaning on the controller in the case of an artificial agent. Therefore, we want to embed these measures into the sensori-motor loop and analyze their behavior in relation to the dynamics of the body and environment. Although the measures are only focusing on the controller, there have been simulated experiments with evolving embodied agents, interacting with their environment, in the context of IIT. In Edlund et al. (2011) the authors measure the integrated information values for simulated evolving artificial agents in a maze and conclude that integrated information grows with the fitness of the agents. Increasing the complexity of the environment leads in Albantakis et al. (2014) to the conclusion that integrated information needs to increase in order to capture a more complex environment. In Albantakis and Tononi (2015) the authors go one step further and conclude from experiments with elementary cellular automata and adaptive logic-gate networks that a high integrated information value increases the likelihood of a rich dynamical behavior. All of these examples focus on the measures in the controller in order to analyze what kind of cause-effect structure makes a difference intrinsically. Since we are interested in an embodied agent solving a task, we want to emphasize the importance of the interaction of the agent's body with the world and additionally measure this interaction explicitly. This leads us to the concept of morphological computation.

Morphological computation is the reduction of computational complexity for the controller resulting from the interaction between the body and the world, as described in Ghazi-Zahedi (2019). There are different ways in which the body can lift the burden of the brain, as discussed in Müller and Hoffmann (2017). An example for morphological computation is the bird using its magneto-reception mentioned earlier in the introduction. Another case of morphological computation would be a human grabbing a fragile object compared to a robotic metal hand. The soft tissue of the human hands allows us to be less precise in the calculation of the pressure that we apply. The robot needs to perform more difficult computations and will therefore most likely have a higher integrated information. Does this mean that our experience of this task is less conscious than the experience of the robot? Here we want to take a step back from the abstract concept of consciousness and instead examine the complexity of the tasks. Even though the interactions are not fully controlled by the human brain, the soft skin of the human hand interacts with the object in a more complicated manner than the robot's hand. In this article we want to analyze how the complexity of solving a task is met by the different information flows among the brain, body and environment. In Lungarella and Sporns (2006), the authors find that the information flow in the agent can be affected by changes in the body's morphology. Examining this phenomenon further we will observe shifts in the importance of the information flows depending on the morphology of the body, which directly changes the complexity of the environment for the agent.

Furthermore, we will define two additional groups of agents. For the agents of the first group all the information has to go through the controller, while the controller has no impact on the action for the agents in the second group. These cases demonstrate once more that the antagonistic behavior of morphological computation and integrated information exists regardless of the behavior of the agents. The results of our experiments are presented in section 3.

2. MATERIALS AND METHODS

2.1. Setting

In order to analyze the information flow of an acting agent, we examine the following simple setting. The agents are idealized models of a two-wheeled robot depicted in **Figure 1A**. Each wheel can spin either fast or slow, hence the agents have four different movements and are unable to stop. If both wheels spin fast, then the agent moves 0.6 units of length and if they both spin slow, then the agent moves 0.2. In case of one fast and one slow wheel the agent makes a turn of approximately 10° with a speed of 0.4. The code of the movement of the agents and a video of 5 agents performing random movements can be found in Langer (2021) . The agent's body consists of a blue circle and a blue line marking the back of the agent, depicted in **Figure 1B**. The two black lines are binary sensors that only detect whether they touch an obstacle or not, without reporting the exact distance to it. If a sensor touches a wall it turns green and if the body of the agent touches a wall it turns red.

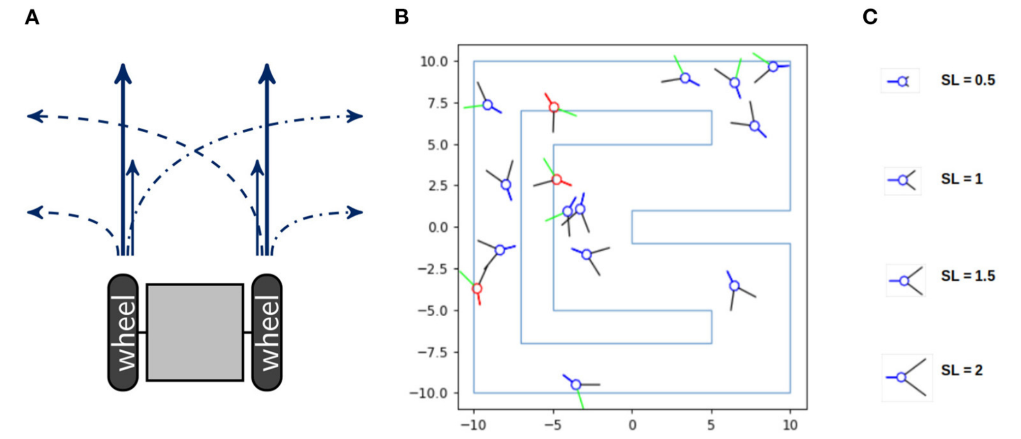


FIGURE 1 | (A) A sketch of a two-wheeled robot and its four different types of movement. **(B)** The racetrack the agents have to survive in and **(C)** the different sensor lengths, named SL, on the right.

Consider a racetrack as shown in **Figure 1B**. The agents die as soon as their bodies touch a wall. Hence the goal for the agents is to stay alive. The design and implementation of the agents and the racetrack is due to Nathaniel Virgo. Although we depicted more than one agent in the environment, these agents do not influence each other.

Additionally, we want to manipulate the amount of potential morphological computation for the agents. There exist different concepts referred to as morphological computation, as thoroughly examined in Müller and Hoffmann (2017), where the authors distinguish between three different categories. These are (1) Morphology facilitating control, (2) Morphology facilitating perception and (3) proper Morphological computation. The notion we will use belongs to the second category and is called "pre-processing" in Ghazi-Zahedi (2019). How well agents perceive their environment can heavily influence the complexity of the task they are facing. One example is the design of the compound eyes of flies, which has been analyzed and used for building an obstacle avoiding robot in Franceschini et al. (1992). Therefore manipulating the qualities of the sensors directly influences the agent's perception and consequently the amount of necessary computation in the controller. Hence changing the length of the sensors influences the agent's ability for interacting with the environment. We will therefore vary the length of the sensors from 0.5 to 2.75. Four different sensor lengths are depicted in **Figure 1C**.

The strategies the agents should use will be calculated by applying the concept of planning as inference as discussed in section 2.2. Utilizing this method we are able to directly determine the optimal behaviors without having to train any agents.

Before we discuss this further, we will first present the control architecture of the agents in the next section.

2.1.1. The Agents

We model the whole system by using the sensori-motor loop as depicted in **Figure 2A**. There the information about the world is received by be the sensors, which send their information to the controller and directly to the actuators. This direct connection between the sensors and the actuators enables the agent to have a response to certain stimuli, without the need for integrating the information in the controller. The controller processes the information from the sensors and also influences the actuators, which in turn have an effect on the world. The sensori-motor loop, also called action-perception circle, has been analyzed and discussed in, for example, Klyubin et al. (2004), Ay and Zahedi (2014), and Ay and Löhr (2015).

Unfolding the connections among the different parts of the agent and its environment for one timestep leads to the depiction in **Figure 2B**. The agents have two sensor S_t^1, S_t^2, two controller C_t^1, C_t^2 and two actuator nodes A_t^1, A_2^2. The sensors and controllers send their information to the actuators and controllers in the next point in time. The sensors are only influenced by the world W and the world is only affected by the actuators and the last world state.

To simplify we only draw one node for each S, A and C in the following graphs.

The behavior of the agents is governed by a probabilistic law, which can be modeled as the following discrete multivariate time-homogeneous Markov process

$$(X_t)_{t\in\mathbb{N}} = (W_t, S_t, A_t, C_t)_{t\in\mathbb{N}}$$

with the state space $\mathcal{X} = \mathcal{W} \times \mathcal{S} \times \mathcal{A} \times \mathcal{C}$ and the distribution

$$P(x_0, \ldots, x_{t+1}) = P(x_0) \prod_{i=1}^{t+1} P(x_t|x_{t-1})$$

$$P(x_{t+1}|x_t) = P(w_{t+1}|w_t, a_t) \prod_k P(s_{t+1}^k|w_{t+1})$$

$$\prod_i P(a_{t+1}^i|s_t, c_t) \prod_j P(c_{t+1}^j|s_t, c_t).$$

The corresponding directed acyclic graph is depicted in **Figure 3A**. See Lauritzen (1996) for more information on the

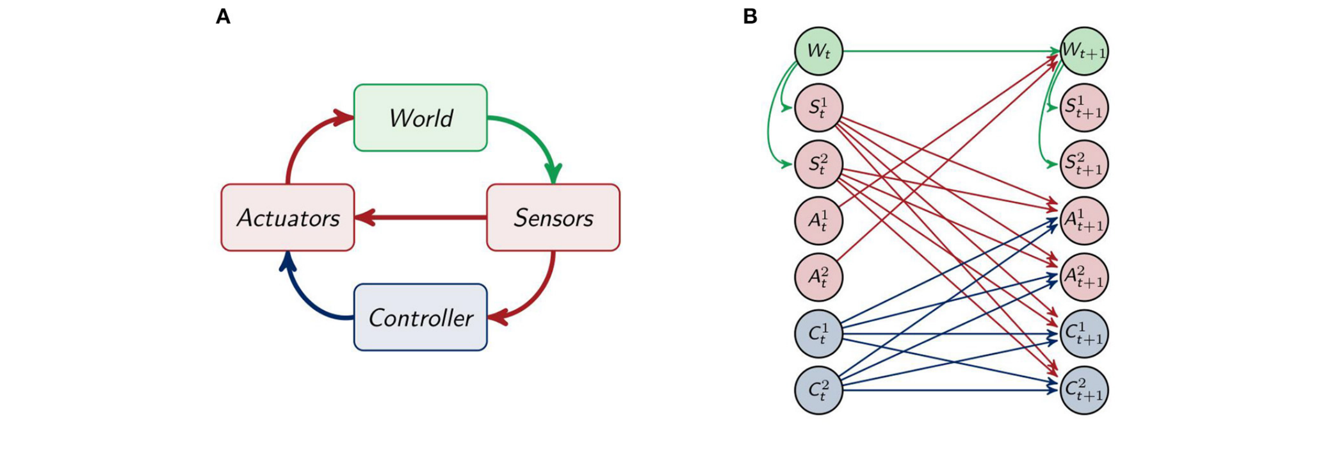


FIGURE 2 | (A) The sensori-motor loop and **(B)** the architecture of the agents.

relationship between graphs and graphical models. Throughout this article we will assume that the distributions on $\mathcal{X}$ are strictly positive.

In the next section we will take a closer look at the role of the environment.

2.1.2. The Environment

The Markov process defined above describes the interactions between the agent and its environment in terms of a joint distribution. Note that the distributions discussed in this section determine the information flow in the system. The optimization of this flow will require a planning process which we are going to address in the next section. Since the agent has only access to the world through the sensors, we replace

$$P(w_{t+1}|w_t, a_t)\prod_k P(s^k_{t+1}|w_{t+1})$$

using information intrinsically known to the agent. In order to do that, we will look closer at one step in time $P(x_t, x_{t+1}) = P(x_t)P(x_{t+1}|x_t)$. Reducing the focus to one step in time means that we need to define an initial distribution that takes into account the past of the agent. In **Figure 3A** we see that the sensors S_t, actuators A_t and controller nodes C_t are conditionally independent given the past, but marginalizing to the point in time t leads to additional connections. More precisely marginalizing to one timestep results in undirected edges between S_t, A_t and C_t. Here we will assume that the environment only influences the sensors, even in the graph marginalized to one timestep as depicted in **Figure 3B**. We will then sum over $w_t, w_{t+1} \in \mathcal{W}$ in order to get a Markov process that only depends on the variables known to the agent.

Proposition 1. *Marginalizing the distribution, that corresponds to the graph (B) in **Figure 3**, that is*

$$P(x_t, x_{t+1}) = P(w_t) \cdot P(s_t, a_t, c_t|w_t) \cdot P(w_{t+1}|w_t, a_t) \prod_k P(s^k_{t+1}|w_{t+1}) \prod_i P(a^i_{t+1}|s_t, c_t) \prod_j P(c^j_{t+1}|s_t, c_t)$$

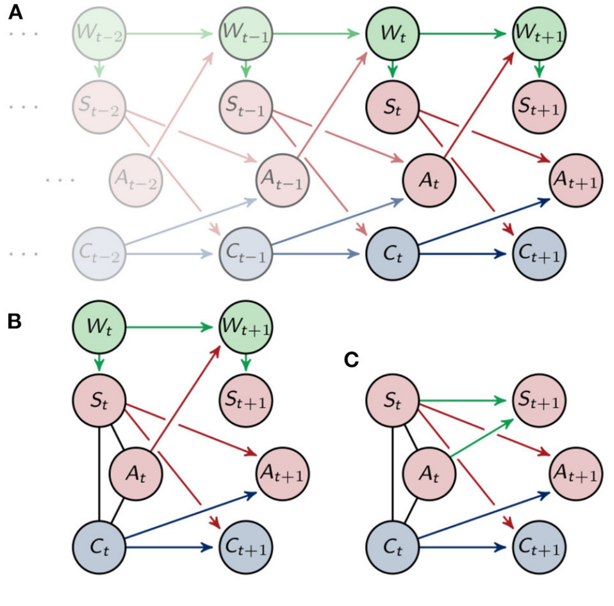


FIGURE 3 | (A) Graphical representation of the Markov process $(X_t)_{t\in\mathcal{N}}$. **(B)** Graphical representation of one timestep and **(C)** the marginalized graph.

over $(w_t, w_{t+1}) \in \mathcal{W} \times \mathcal{W}$ leads to the following Markov process

$$P(s_t, a_t, c_t, s_{t+1}, a_{t+1}, c_{t+1}) = P(s_t, a_t, c_t) \cdot \prod_i P(a^i_{t+1}|s_t, c_t) \prod_j P(c^j_{t+1}|s_t, c_t) \cdot P(s_{t+1}|s_t, a_t).$$

The proof can be found in the **Supplementary Material**.

The new process describes the behavior of the environment with information known to the agent and is shown in **Figure 3C**. A similar distribution is also used in Ghazi-Zahedi and Ay (2013) in section 3.3.1. and in Ghazi-Zahedi (2019). There it is derived

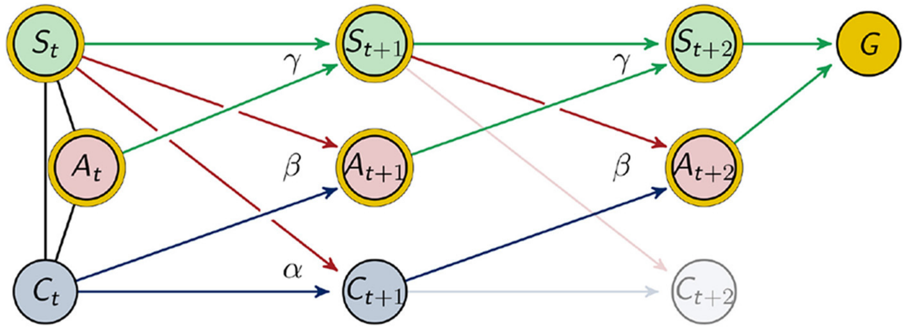


FIGURE 4 | Graphical representation of two timesteps.

by taking $P(S_{t+1}|S_t)$ as the intrinsically available information of $P(W_{t+1}|W_t)$.

We sample this distribution $\tilde{P}(S_{t+1}, S_t, A_t)$ for every sensor length, by storing 20.000.000 sensor and motor values for agents starting in a random place in the arena, performing arbitrary movements. We denote all the sampled and therefore fixed distributions by $\tilde{P}$.

Since we are now able to define a set of distributions that describe the interaction between the agent and the world according to the sensori-motor loop, we will present the method to find the optimal behavior in the next section.

2.2. Optimizing the Behavior

In order to calculate the optimal behavior of the agents, we will use the concept of planning as inference. This was originally proposed by Attias in Attias (2003) and further developed by Toussaint and collegues in Toussaint et al. (2006), Toussaint (2009), and Toussaint et al. (2008) as a theory of planning under uncertainty. There the conditional distribution describing the action of the agent is considered to be a hidden variable that has to be optimized. This is done by using the EM-algorithm, which is equivalent to the information theoretic em-algorithm in this case. We use the em-algorithm, because of its intuitive geometric nature. More details can be found in the **Supplementary Material**. The em-algorithm is well known and was proposed in Csiszár and Tsunády (1984), further discussed in Amari (1995) and Amari et al. (1992). The resulting distribution maximizes the likelihood of achieving the predefined goal, but might be a local optimum depending on the initial distribution. Normally this is a disadvantage, but in our setting it allows us to analyze various strategies by using different initial distributions.

The goal of the agents in our example is to maximize the probability of being alive after the next two movements. To make at least two steps is necessary since we want the connection between C_t and C_{t+1} to have an impact on the outcome. This can be seen in **Figure 4**.

We will denote the goal variable by G with the state space $\mathcal{G} = \{0, 1\}$, where $P(g_1) := P(g = 1)$ refers to the probability of the agent to be alive. Since the agent moves twice, this distribution depends on the states of the last three sensor and motor states

$$\tilde{P}(G|S_{t+2}, S_{t+1}, S_t, A_{t+2}, A_{t+1}, A_t).$$

The variable G depends on the nodes that are marked with a golden circle in **Figure 4**. We sampled this distribution for every sensor length, as described in the previous section in the context of $\tilde{P}(S_{t+1}, S_t, A_t)$.

The architecture of the agents considered in this article was discussed in the last sections. There we outlined how we sample the distribution $\gamma = \tilde{P}(S_{t+1}|S_t, A_t)$ that describes the influence the agent has on itself through the world. The distributions influencing the behavior of the agents are

$$\beta = P(A_{t+1}|S_t, C_t) \qquad \text{and} \qquad \alpha = P(C_{t+1}|S_t, C_t).$$

Hence we will treat (A_{t+1}, C_{t+1}) as hidden variables and optimize their distributions with respect to the goal. We denote these distributions by α, β and γ in order to emphasize that the process is time-homogeneous, meaning that $P(A_{t+1}|S_t, C_t) = P(A_{t+2}|S_{t+1}, C_{t+1})$, $P(S_{t+1}|S_t, A_t) = P(S_{t+2}|S_{t+1}, A_{t+1})$ and $P(C_{t+1}|S_t, C_t) = P(C_{t+2}|S_{t+1}, C_{t+1})$ as indicated in **Figure 4**. Note that the above mentioned homogeneity does not imply stationarity.

It remains to define the initial distribution $P(S_t, C_t, A_t)$. In the original planning as inference framework an action sequence is selected conditioned on the final goal state and an initial observation, as described in Attias (2003). Here, we do not want to restrict the agents to an initial observation S_t. Instead we first write the initial distribution in the following form

$$P(s_t, c_t, a_t) = P(c_t|a_t, s_t)P(s_t|a_t)P(a_t).$$

Using the sampled distribution $\tilde{P}(S_{t+1}, S_t, A_t)$, we are able to calculate $\tilde{P}(S_t|A_t)$ and set $P(s_t|a_t) = \tilde{P}(s_t|a_t)$. The remaining distributions $P(c_t|a_t, s_t)$ and $P(a_t)$ are also treated as variables and optimized using the em-algorithm. This approach leads to the optimal starting conditions for the agents. The details of the optimization are described in the **Supplementary Material**.

2.3. Measures of the Information Flow

In this section we will define the different measures. These are information theoretic measures that use the KL-divergence to calculate the difference between the original distribution and a split distribution. This split distribution is the one that is closest to the original distribution without having the connection that we want to measure.

Definition 1 (Measure Ψ). *Let $M \subset P^{\circ}(\mathcal{Z})$ be a set of probability distributions corresponding to a split system. Then we define the measure Ψ, by minimizing the KL-divergence between M and the full distribution P to quantify the strength of the connections missing in the split system*

$$\Psi = \inf_{Q \in M} D(P \parallel Q) = \sum_{z} P(z) \, log \, \frac{P(z)}{Q(z)}.$$

Note that this measure depends on M, the set of split distributions.

Every discussed measure has a closed form solution and can be written in the form of sums of conditional mutual information terms.

Definition 2 (Conditional Mutual Information). *Let (Z_1, Z_2, Z_3) be a random vector on $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$ with the distribution P. The conditional mutual information of the random variables Z_1 and Z_2 given Z_3 is defined as*

$$\begin{aligned} I(Z_1; Z_2|Z_3) &= \sum_{z_1 \in \mathcal{Z}_1} \sum_{z_2 \in \mathcal{Z}_2} \sum_{z_3 \in \mathcal{Z}_3} P(z_1, z_2, z_3) \, log \left(\frac{P(z_1, z_2|z_3)}{P(z_1|z_3)P(z_2|z_3)} \right) \\ &= \sum_{z_1 \in \mathcal{Z}_1} \sum_{z_2 \in \mathcal{Z}_2} \sum_{z_3 \in \mathcal{Z}_3} P(z_1, z_2, z_3) \, log \left(\frac{P(z_1|z_2, z_3)}{P(z_1|z_3)} \right). \end{aligned}$$

If $I(Z_1; Z_2|Z_3) = 0$, then Z_1 is independent of Z_2 given Z_3. Therefore, this quantifies the connection between Z_1 and Z_2, while Z_3 is fixed. Additionally, we emphasize this by marking the respective connection quantified by the measure in a graph as a dashed connection. To simplify the figures we only depict one timestep, but the connections between (Y_{t+1}, Y_{t+2}) are the same as the connections between (Y_t, Y_{t+1}).

The base of the logarithms in the definitions above is 2, hence the unit of all the measures defined below is bits.

Although these measures were originally defined for only one timestep, we will introduce them directly tailored to our setting with two timesteps.

2.3.1. Integrated Information and Morphological Computation

The two measures discussed in this section each quantify the information flow among the same type of node in different points in time. Integrated information only considers the nodes inside the controller and therefore measures the information flow inside the agent, while morphological computation is concerned with the exterior perspective and measures the information flow between the sensors.

2.3.1.1. Integrated Information

The measure Φ_T restricts itself to the controller nodes and can be seen in the context of the Integrated Information Theory of consciousness (Tononi, 2004). This theory was discussed in the introduction. It aims at measuring the strength of the connections among different nodes across different points in time, in other words, the connections that integrate the information. Since every influence on C_{t+1} is known in our setting, we are able to use the measure Φ_T proposed in Langer and Ay (2020). This measure is defined in the following way

$$\Phi_T = \sum_{\tau \in \{t, t+1\}} \sum_{j} I(C^j_{\tau+1}; C^{I \setminus \{j\}}_{\tau} | C^j_{\tau}, S_{\tau})$$

and depicted as (b) in **Figure 5**. In the definition above, $I(C^j_{t+1}; C^{I\setminus\{j\}}_t | C^j_t, S_t)$ denotes the conditional mutual information, described in Definition 2, and $I \setminus \{j\}$ is the set of indices of controller nodes without j. For two controller nodes and $j = 2$ this would be $\{1, 2\} \setminus \{2\} = \{1\}$. Hence Φ_T measures the connections between C^i_t and C^j_{j+1} with $i, j \in \{1, 2\}$ and $i \neq j$.

A proof of the closed form solution can be found in Langer and Ay (2020). All the following measures can be proven in a similar way.

2.3.1.2. Morphological Computation

In Ghazi-Zahedi (2019) morphological computation was referred to as morphological intelligence and characterized in Definition 1.1. as follows

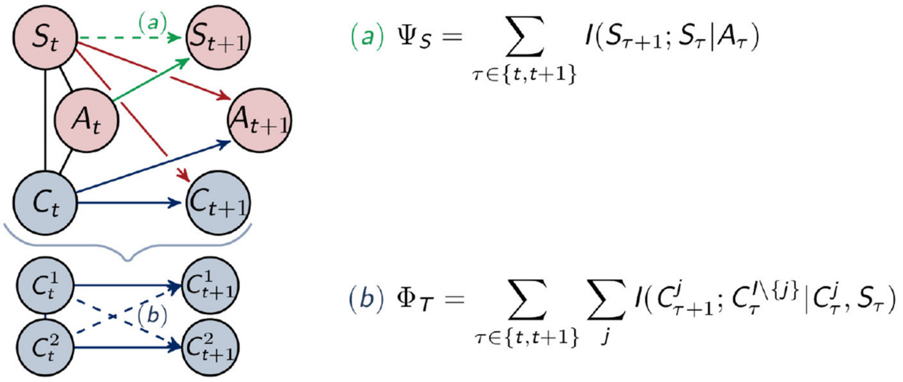
> "Morphological Intelligence is the reduction of computational cost for the brain (or controller) resulting from the exploitation of the morphology and its interaction with the environment."

There exists a variety of measures for morphological computation, described for example in Ghazi-Zahedi (2019) and Ghazi-Zahedi et al. (2017). The distribution $\tilde{P}(S_{t+1}|S_t, A_t)$ describes the influence the agent has on itself through the environment. Hence this distribution is dependent on the environment and the morphology of the agent. The interplay between environment and body is influenced by the length of the sensors.

In Ghazi-Zahedi and Ay (2013) the authors define the following measure for morphological computation, which depends on $\tilde{P}(S_{t+1}|S_t, A_t)$. It quantifies the strength of the influence of the past sensory input on the next sensory input given the last action as

$$\Psi_S = \sum_{\tau \in \{t, t+1\}} I(S_{\tau+1}; S_{\tau}|A_{\tau})$$

which corresponds to $ASOC_W$ defined in Ghazi-Zahedi (2019) in Definition 3.1.3. There the author compares the different measures numerically and concludes in the chapter 4.9 that the measure following the approach of Ψ_S, but defined directly on the world states, has advantages over other formulations and is therefore the recommended one. We will follow this reasoning and consider Ψ_S to be the measure of morphological computation.



$$(a)\ \Psi_S = \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; S_\tau | A_\tau)$$

$$(b)\ \Phi_T = \sum_{\tau\in\{t,t+1\}} \sum_{j} I(C^j_{\tau+1}; C^{I\setminus\{j\}}_\tau | C^j_\tau, S_\tau)$$

FIGURE 5 | Calculation of the measures for morphological computation (a) and integrated information (b).

2.3.2. Measures for Information Flows Between Different Types of Nodes

We will observe that the measures for integrated information and morphological computation behave antagonistically. This, however, does not lead to a definitive conclusion about how much of the behavior of the agent is determined by the controller. Intuitively, it might be the case, that the agent acts regardless of all the information integrated in the controller. In order to understand the influences leading to the actions of the agents, we will present four additional measures for the four remaining connections in the graph and a measure quantifying the total information flow. These are depicted in **Figure 6**.

2.3.2.1. Reactive Control

Reactive control describes a direct stimuli response, meaning that the sensors send their unprocessed information directly to the actuators. We are measuring this by the value Ψ_R. The corresponding split distribution results from removing the connection between S_t and A_{t+1}

$$\Psi_R = \sum_{\tau\in\{t,t+1\}} \sum_{i} I(A^i_{\tau+1}; S_\tau | C_\tau).$$

2.3.2.2. Action Effect

We are able to quantify the effect of the action on the next sensory state by calculating

$$\Psi_A = \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; A_\tau | S_\tau).$$

This measures the amount of control an agent has. Hence in Ghazi-Zahedi and Ay (2013) this measure was normalized and inverted in order to quantify morphological computation. The differences between this approach and Ψ_S are further discussed in section 4.9 in Ghazi-Zahedi (2019).

2.3.2.3. Sensory Information

The commands the controller sends to the actuators should be based on the information received from the sensors. Therefore, we will additionally calculate the strength of the information flow from the sensor to the controller nodes. The smaller this value is, the more likely it is that the controller converged to a general strategy and performs this blindly without including the information from the sensors. We will call this "sensory information," Ψ_{SI},

$$\Psi_{SI} = \sum_{\tau\in\{t,t+1\}} \sum_{j} I(C^j_{\tau+1}; S_\tau | C_\tau).$$

2.3.2.4. Control

Since we are looking at an embodied agent, we additionally want to measure how much of the information processed in the controller has an actual impact on the behavior of the agent. We will term the measure quantifying the strength of the impact of the controller on the actuators "control," Ψ_C,

$$\Psi_C = \sum_{\tau\in\{t,t+1\}} \sum_{i} I(A^i_{\tau+1}; C_\tau | S_\tau).$$

2.3.2.5. Total Information Flow

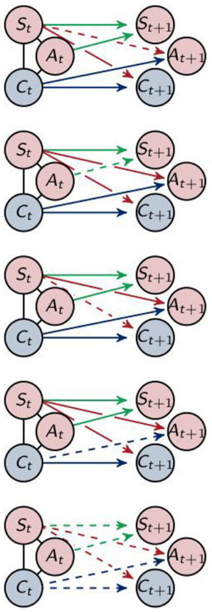
The last measure quantifies the total information flow, Ψ_{TIF}. In this case two points in time are independent of each other in the split system, as depicted in **Figure 6**,

$$\begin{aligned}\Psi_{TIF} = & \sum_{\tau\in\{t,t+1\}} \sum_{i,j} I(S_{\tau+1}; S_\tau, A_\tau) + I(A^i_{\tau+1}; S_\tau, C_\tau) \\ & + I(C^j_{\tau+1}; C_\tau, S_\tau).\end{aligned}$$

The total information flow is an upper bound for all the other measures defined in the previous sections.

3. RESULTS

In this section we will present the results of our experiments. The length of the sensors are varied from 0.5 to 2.75 in steps of 0.25. We took 100 random input distributions $\bar{P}$. Each time the algorithm takes at least 1,000 iteration steps and stops when the difference between the likelihood of the goal is smaller than $1 * 10^{-5}$.



(c) $\Psi_R = \sum_{\tau\in\{t,t+1\}} \sum_i I(A^i_{\tau+1}; S_\tau | C_\tau)$

(d) $\Psi_A = \sum_{\tau\in\{t,t+1\}} I(S_{\tau+1}; A_\tau | S_\tau)$

(e) $\Psi_{SI} = \sum_{\tau\in\{t,t+1\}} \sum_j I(C^j_{\tau+1}; S_\tau | C_\tau)$

(f) $\Psi_C = \sum_{\tau\in\{t,t+1\}} \sum_i I(A^i_{\tau+1}; C_\tau | S_\tau)$

(g) $\Psi_{TIF} = \sum_{\tau\in\{t,t+1\}} \sum_{i,j} I(S_{\tau+1}; S_\tau, A_\tau) + I(A^i_{\tau+1}; S_\tau, C_\tau) + I(C^j_{\tau+1}; C_\tau, S_\tau)$

FIGURE 6 | Calculation of the measures for (c) reactive control, (d) action effect, (e) sensory information, (f) control and (g) total information flow.

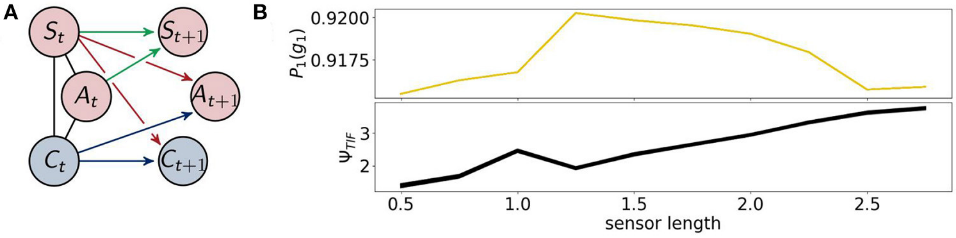


FIGURE 7 | (A) The architecture of the fully coupled agents and **(B)** the probability of survival (top) and the total information flow Ψ_{TIF} (bottom).

3.1. Fully Coupled Agents

The architecture of the fully coupled agents are the ones described in section 2.1.1 as shown in **Figure 7** on the left. We will refer to the optimized distribution of a fully coupled agent by P_1, hence $P_1(g_1)$ is the probability with which the agents survive. This value is depicted in **Figure 7** on the top right. The agents perform best between a sensor length of 1.25 and 2.25. If the sensors are too long or too short their information is not useful to assure the survival of the agents.

The total information flow Ψ_{TIF} in **Figure 7** on the bottom right exhibits an almost monotonic increase, except for a local maximum at a sensor length of 1. We will discuss this sensor length below in the context of Ψ_R and Ψ_A.

Now we are going to present the results for integrated information Φ_T and morphological computation Ψ_S, depicted in **Figure 8**.

We observe that Φ_T monotonically decreases as the sensors become larger. Directly to the right, Ψ_S exhibits the opposite dynamic. It quantifies the influence of the past sensory input on the next sensory input given the action. Hence, taking the perspective of the agent, Ψ_S describes the extrinsic information flow, whereas Φ_T only depends on the controller nodes and quantifies therefore, the most intrinsic information flow. So these measures exhibit an antagonistic relationship between the outside and the inside, meaning between morphological computation and integrated information.

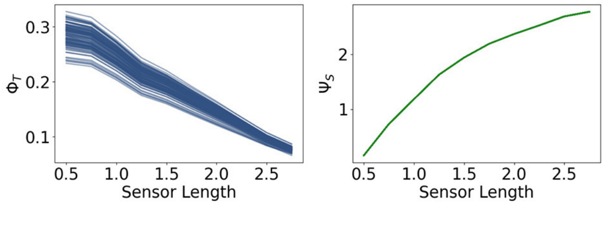


FIGURE 8 | Integrated information Φ_T and morphological computation Ψ_S for the fully coupled agents.

Note that the total information flow, Ψ_{TIF}, is a sum of three mutual information terms and that the first term $I(S_{t+1}; S_t, A_t)$ is an upper bound of Ψ_S, the measure for morphological computation. Since Ψ_S is particularly high compared to the other measures, the dynamics of $I(S_{t+1}; S_t, A_t)$ are dominating Ψ_{TIF}, leading to the monotonic increase in **Figure 7**.

In **Figure 9** in the first row we see the measures Ψ_{SI} and Ψ_C. The measure Ψ_{SI} quantifies how important the information flow from the sensors to the controller is. For a length below 1 the sensors are too short and above approximately 2 too long to carry information that is valuable for the controller. The importance of the commands sent from the controller to the actuators is measured by Ψ_C. Between 0.5 and 1.25 this value is very close to 0, which means, that the controller has next to no influence on the behavior of the agent. In this case the sensors are so short that the agents need to react directly to it.

Hence, although Φ_T has its maximum values at a sensor length of 0.5, the integrated information does not have a significant impact on the behavior of the agents. Therefore the importance of the information flow in the controller of an embodied agent depends additionally on the information flowing to and from the controller.

The measure for reactive control is shown in the second row. In the case of short sensors, the information needs to get passed directly to the actuators. Now we will compare Ψ_R to Ψ_A, depicted on the bottom right in **Figure 9**. The latter one is defined as the action effect, meaning the higher Ψ_A is, the more influence the actuators have on the next sensor state. The maximum of Ψ_R and Ψ_A are at a sensor length of 1, which results in the local maximum of Ψ_{TIF} in **Figure 7**. Both graphs show similar dynamics between sensors of length 1 to 2.25. If the sensors are too small, the information needs to pass directly to the actuators, but the actuators might not be able to assure survival and therefore Ψ_R is high, while Ψ_A is low. In the case of very long sensors, they detect a wall with a high probability, so that the next sensory state will again detect a wall regardless of the action taken. This leads to a high Ψ_R and a low Ψ_A.

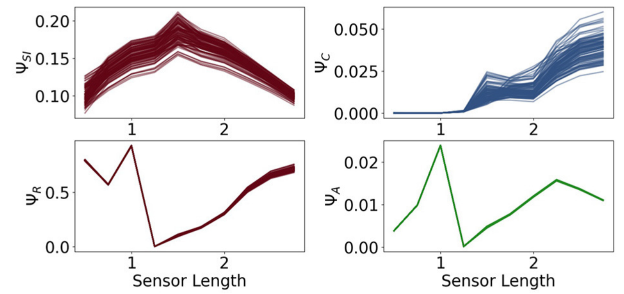


FIGURE 9 | The measures for control Ψ_C, sensory information Ψ_{SI}, reactive control Ψ_R and action effect Ψ_A for the fully coupled agents.

At a sensor length of 1.25, Ψ_R is close to 0, as well as Ψ_C and Ψ_A, which suggests that the algorithm converged to an optimum in which the next sensor state is not dependent on the action and the action is not dependent on the last sensor state.

At a first glance the values of Ψ_A and Ψ_C seem to be insignificant compared to the other measures, but note that the relatively small amount is an expected result in these experiments. The last sensor state has a very high influence on the next sensor state and on the next action, since an agent that is not touching a wall will most likely not touch a wall in the next step and move slowly, whereas an agent touching a wall will steer away and, depending on the length of the sensors, probably touch a wall in the next step. Nevertheless, if Ψ_A and Ψ_C are not zero, then there exists an information flow and therefore an influence from the actuators to the sensors and from the controller nodes to the actuators. Hence observing the dynamics and relating them to the other measures does lead to insights to the interplay of the different information flows.

In order to further substantiate the results of our analysis, we will now examine two subclasses of agents. We will directly manipulating the architecture of the agents so that the influence on the actuators are limited. Hence we will gain insights on the importance of reactive control and the controller for the behavior of the agent. The first subclass contains agents that are incapable of reactive control and therefore all the information has to flow through the controller. Hence we call them controller driven agents in section 3.2. The second class consists of agents in which the controller has no impact on the actuators. These will be called reactive control agents and discussed in section 3.3.

3.2. Controller Driven Agents

Now, we will discuss the results for the agents that are not able to use reactive control. These are displayed in **Figure 10** on the left. We will refer to the optimized distributions by P_2.

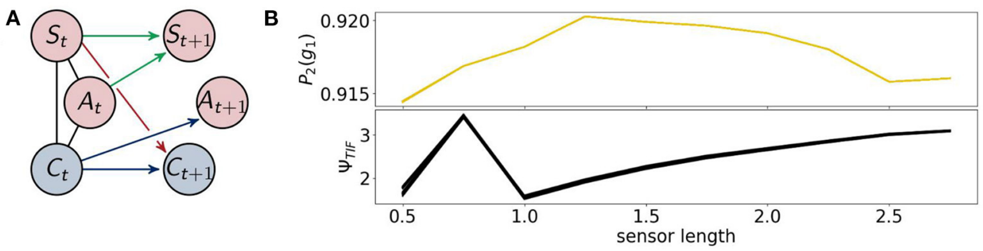


FIGURE 10 | (A) the architecture of the controller driven agents and **(B)** the probability of survival (top) and the total information flow Ψ_{TIF}(bottom).

Note that these agents are a subclass of the fully coupled ones. Hence optimizing the likelihood of success for these agents should not lead to a higher value for success than for the fully coupled agents. Since we are using the em-algorithm that converges to local minima, however, we observe that controller driven agents have a higher probability of success around a sensor length 1, as depicted on the right in **Figure 10**.

The results of the total information flow are similar compared to the case of the fully coupled agents after a sensor length of 1. In this case Ψ_{TIF} has a global maximum at 0.75, which we will discuss in the context of Ψ_{SI} and Ψ_A.

The measures Φ_T and Ψ_S show in **Figure 11** approximately the same values as in **Figure 8**. There is no change in the dynamics of Ψ_S, but Φ_T is lower than before at a sensor length of 0.5. Note that Ψ_C in **Figure 12** is significantly higher in this case, so that the integrated information makes an impact on the actuators.

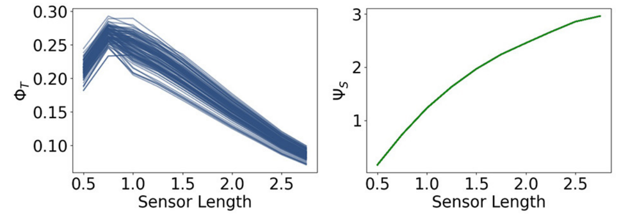


FIGURE 11 | Integrated information and morphological computation for the controller driven agents.

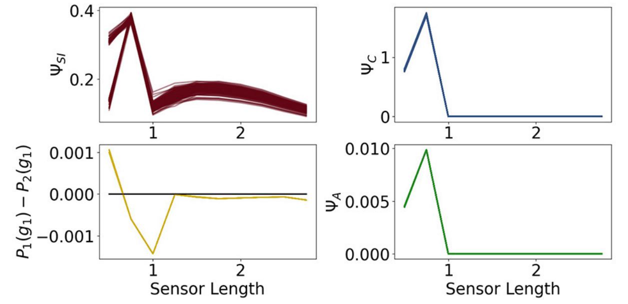


FIGURE 12 | The measures for control Ψ_C, sensory information Ψ_{SI}, action effect Ψ_A for the controller driven agents and the performance difference the fully coupled agents and the reactive ones.

All of the measures corresponding to the controller have a spike at 0.75, at which point these agents perform better than the ones with the ability for reactive control as can be seen in the graph on the bottom left of **Figure 12**. There the total information flow Ψ_{TIF}, depicted in **Figure 10**, reaches its maximum. This spike can also be observed in Ψ_A, meaning that the influence of the actuators on the next sensory input given the last sensory input is high.

Additionally, looking at the goal difference depicted on the bottom left in **Figure 12**, we see that these agents perform better than the fully coupled agents for the sensors being longer than 0.5. The black line marks the value 0. After a sensor length of 1 the measures Ψ_C and Ψ_A and show that the information flows from the controller to the actuators and from the actuators to the sensors are barely existent. Therefore, we come to the conclusion, that the agents converged to an optimum in which the actuators do not depend on the sensory input and have no influence on the next sensory state. Note that Φ_T still shows the decreasing behavior, even though it has no impact on the actions of the agent.

3.3. Reactive Control Agents

The architecture of the reactive control agents is shown in **Figure 13** on the left. Here the controller has no influence on the actuators. On the right we see the probability of survival $P_3(g_1)$.

There is now significant difference between the total information flow of the fully coupled agents and the total information flow in this case.

The measures Φ_T and Ψ_S show in **Figure 14** the same antagonistic behavior as in the fully coupled case. This

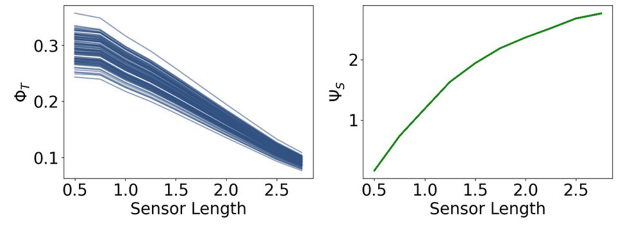


FIGURE 14 | Integrated information and morphological computation for the reactive control agents.

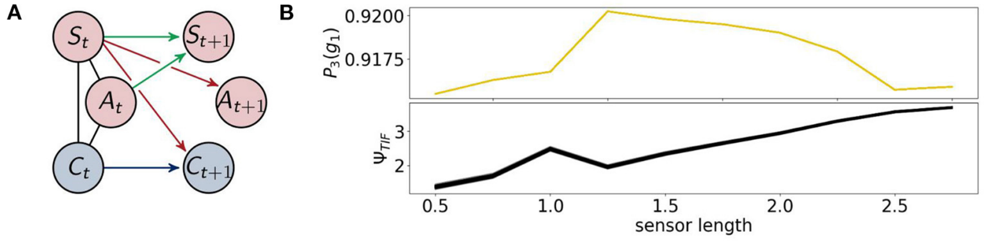


FIGURE 13 | (A) The architecture of the reactive control agents, **(B)** the probability of survival (top) and the total information flow Ψ_{TIF} (bottom).

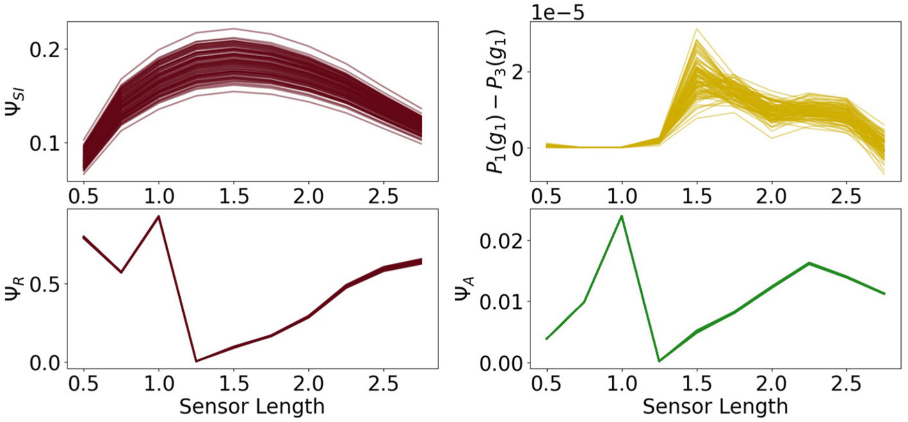


FIGURE 15 | The measures for control morphological computation Ψ_S, reactive control Ψ_R and action effect Ψ_A for the reactive control agents and the performance difference between the fully coupled agents and the reactive ones.

demonstrates once more that only using integrated information as a measure in the case of embodied agents does not suffice if we want to understand the agent's behavior.

A closer examination of the difference in performance, depicted on the top right in **Figure 15**, reveals that the agents connected to a controller perform better for sensors between 1.25 and 2.5. Looking back at **Figure 9**, we see that this is approximately the region in which Ψ_C and Ψ_{SI} are both high. This supports the idea that integrated information has an impact on the behavior, when at the same time the information flows to and from the controller are high.

The other measures show the same dynamics as the corresponding measures for the fully coupled agents.

4. DISCUSSION

In this article we combine different techniques in order to create a framework to analyze the information flow among an agents body, its controller and the environment. The main question we want to approach is how the complexity of solving a task is distributed among these different interacting parts.

We demonstrate the steps in the analysis with the example of small simulated agents that are not allowed to touch the walls of a racetrack. These agents have a sufficiently simple architecture such that we are able to rigorously analyze the different information flows. Additionally, we can examine the dynamics of the information theoretic measures of an agent under changing morphological circumstances by modifying the length of the sensors.

We calculate the optimal behavior by using the concept of planning as inference which allows us to model the conditional distributions determining the actions of the agents as latent variables. Using the information geometric em-algorithm, we are able to optimize the latent variables such that the probability of success is maximal. Here, the expectation maximization EM algorithm used in statistics is equivalent to the em-algorithm, but we chose to present the em-algorithm, because it has an intuitive geometric interpretation. This algorithm is guaranteed to converge, but converges to different (local) optima depending on the starting distribution. Hence this allows us to analyze various kinds of strategies that lead to a reasonably successful agent.

The distributions that are optimal regarding reaching a goal are then analyzed by applying seven information theoretic measures. We use the measure Φ_T to calculate the integrated information in the controller and we demonstrate that, although the agents have goal optimized policies, this value can be high even in cases in which it has no behavioral relevance. Therefore, the importance of the information flow in the controller of an embodied agent additionally depends on the information flow to and from the controller, measured by Ψ_{SI} and Ψ_C. Hence, if we want to fully understand the impact integrated information has on the behavior of an agent, it is not sufficient to only calculate an integrated information measure. This is supported by the comparison of the fully coupled agents to the reactive ones, the agents in which the controller has no impact on the actuators. It shows that the controller has a positive influence on the performance of the agents exactly in the cases in which Ψ_{SI} and Ψ_C are both high.

Comparing the morphological computation, measured by Ψ_S, to the integrated information reveals an antagonistic relationship between them. The more the agent's body interacts with its environment, the less information is integrated.

The measure for reactive control Ψ_R displays a dynamic similar to the action effect Ψ_A. Removing the ability to send information from the sensors directly to the actuators, in the controller driven agents, leads to agents that perform an action regardless of the sensor input for a sensor length greater than 1.

Finally, the total information flow is an upper bound for the other measures. Therefore Ψ_{TIF} combined with the measures above give us a notion of which information flow has the most influence on the system.

All in all, we present a method to completely examine the information flow among the controller, body and environment of an agent. This gives us insights into how the complexity of the task is met by the different interacting components. We observe how the morphology of the body and the architecture of the agents influence the internal information flows. The example discussed in this article is limited by its simplicity, but even in this scenario, we were able to demonstrate the value of examining the different measures.

We will continue to develop these concepts further to be able to efficiently analyze more complicated agents and tasks and test them on humanoid robots. A humanoid robot can perform for example a reaching movement, which is a goal directed task that allows for more degrees of freedom and the need to integrate different information sources such as visual information and the angle of the joints.

Furthermore, we have seen in the examples presented in this paper, that some tasks can be performed without involvement of the controller. In contrast to the agents in this article, which are optimized directly using planning as inference, natural agents learn to control their body and to interact with their environment gradually. It is intuitive to assume that learning a new task requires much more computation in the controller than executing an already acquired skill. Hence, it is important to analyze the temporal dynamics of the integrated information and morphological computation measures during the learning process to gain insights into potential learning phases. These different learning phases may lead us one step closer to understanding the emergence of the senses of agency and body ownership, two concepts closely related to the human minimal self (Gallagher, 2000).

Using an agent with a more complicated morphology can lead to the opportunity to study the "degrees of freedom" problem, formulated in motor control theory. In his influential work (Bernstein, 1967) addresses the difficulties resulting from the many degrees of freedom within a human body, namely the problem of choosing a particular motor action out of a number of options that lead to the same outcome. In Bernstein (1967), in the chapter "Conclusions toward the study of motor co-ordination," he makes the following observation:

> "All these many sources of indeterminacy lead to the same end result; which is that the *motor effect of a central impulse cannot be decided at the centre* but is decided entirely at the periphery: at the last spinal and myoneural synapse, at the muscle, in the mechanical and anatomical change of forces in the limb being moved, etc."

He thus emphasizes the importance of the morphology of the body for the actual movement.

There have been a number of theories further discussing this topic. In Todorov and Jordan (2002), for example, the authors propose a computational level theory based on stochastic optimal feedback control. The resulting "minimum intervention principle" highlights the importance of variability in task-irrelevant dimensions. It would be interesting to analyze, whether we observe spikes in the control value and the integrated information that indicate a correctional motor action only for the task-relevant dimensions.

Another theory approaching the degrees of freedom problem is the "equilibrium point hypothesis" by Feldman and colleagues, Asatrian and Feldman (1965) and Feldman (1986). There the control is modeled by shifting equilibrium points in opposing muscles. The usage of the properties of the body in order to achieve co-ordination is directly related to the concept of morphological computation. The authors of Montúfar et al. (2015) study how relatively simple controllers can achieve a set of desired movements through embodiment constraints and call this concept "cheap control."

By applying our framework to more complex tasks, we would expect results agreeing with the observations in Montúfar et al. (2015). Fewer degrees of freedom, which are associated with strong embodiment constraints, should lead to high morphological computation and therefore, following the reasoning of this paper, to a small integrated information value.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/CarlottaLanger/MorphologyShapesIntegratedInformation.

AUTHOR CONTRIBUTIONS

NA and CL: conceptualization and methodology. CL: software, investigation, and writing. NA: supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

The authors acknowledge funding by Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134). Publishing fees supported by Funding Programme *Open Access Publishing* of Hamburg University of Technology (TUHH).

ACKNOWLEDGMENTS

We thank Nathaniel Virgo for the design and implementation of the racetrack and for the introduction to the theory of Planning as Inference.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.716433/full#supplementary-material

REFERENCES

Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput. Biol*. 10:e1003966. doi: 10.1371/journal.pcbi.1003966

Albantakis, L., and Tononi, G. (2015). The intrinsic cause-effect power of discrete dynamical systems—from elementary cellular automata to adapting animats. *Entropy* 17, 5472–5502. doi: 10.3390/e17085472

Amari, S.-I. (1995). Information geometry of the EM and em algorithms for neural networks. *Neural Netw*. 8, 1379–1408. doi: 10.1016/0893-6080(95)00003-8

Amari, S.-I., Kurata, K., and Nagaoka, H. (1992). Information geometry of boltzmann machines. *IEEE Trans. Neural Netw*. 3, 260–271. doi: 10.1109/72.125867

Asatrian, D., and Feldman, A. (1965). On the functional structure of the nervous system during movement control or preservation of a stationary posture. i. mechanographic analysis of the action of a joint during the performance of a postural task. *Biofizika* 10, 837–846.

Attias, H. (2003). "Planning by probabilistic inference," in *Proceeding of the 9th International Workshop on Artificial Intelligence and Statistics, Volume R4 of Proceedings of Machine Learning Research*, eds C. M. Bishop and B. J. Frey (Key West: PMLR), 9–16.

Ay, N., and Löhr, W. (2015). The umwelt of an embodied agent-a measure-theoretic definition. *Theory Biosci*. 134, 105–116. doi: 10.1007/s12064-015-0217-3

Ay, N., and Zahedi, K. (2014). *On the Causal Structure of the Sensorimotor Loop, chapter 9*. Berlin; Heidelberg: Springer Berlin Heidelberg.

Bernstein, N. (1967). *The Co-ordination and Regulation of Movements*. Oxford: Pergamon Press.

Csiszár, I., and Tsunády, G. (1984). "Information geometry and alternating minimization procedures," in *Statistics and Decisions (Supplementary Issue, No.1)*, ed E. F. Dedewicz (Munich: Oldenburg Verlag), 205–237.

Edlund, J., Chaumont, N., Hintze, A., Koch, C., Tononi, G., and Adami, C. (2011). Integrated information increases with fitness in the evolution of animats. *PLoS Comput. Biol*. 7:e1002236. doi: 10.1371/journal.pcbi.1002236

Feldman, A. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *J. Mot. Behav*. 18, 17–54. doi: 10.1080/00222895.1986.10735369

Franceschini, N., Pichon, J.-M., and Blanes, C. (1992). From insect vision to robot vision. *Philos. Trans. R. Soc. B Biol. Sci*. 337, 283–294. doi: 10.1098/rstb.1992.0106

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci*. 4, 14–21. doi: 10.1016/S1364-6613(99)01417-5

Gallagher, S. (2005). *How the Body Shapes the Mind*. New York, NY: Oxford University Press UK.

Ghazi-Zahedi, K. (2019). *Morphological Intelligence*. Cham: Springer.

Ghazi-Zahedi, K., and Ay, N. (2013). Quantifying morphological computation. *Entropy* 15, 1887–1915. doi: 10.3390/e15051887

Ghazi-Zahedi, K., Langer, C., and Ay, N. (2017). Morphological computation: synergy of body and brain. *Entropy* 19:456. doi: 10.3390/e19090456

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004). "Tracking information flow through the environment: simple cases of stigmergy," in *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, ed J. Pollack (Boston, MA: MIT Press), 563–568.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2007). Representations of space and time in the maximization of information flow in the perception-action loop. *Neural Comput*. 19, 2387–2432. doi: 10.1162/neco.2007.19.9.2387

Langer, C. (2021). *Morphology Shapes Integrated Information Github Repository*. Available online at: https://github.com/CarlottaLanger/MorphologyShapesIntegratedInformation.

Langer, C., and Ay, N. (2020). Complexity as causal information integration. *Entropy* 22:1107. doi: 10.3390/e22101107

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Lungarella, M., Pegors, T., Bulwinkle, D., and Sporns, O. (2005). Methods for quantifying the informational structure of sensory and motor data. *Neuroinformatics* 3, 243–262. doi: 10.1385/NI:3,3:243

Lungarella, M., and Sporns, O. (2006). Mapping information flow in sensorimotor networks. *PLoS Comput. Biol*. 2:e144. doi: 10.1371/journal.pcbi.0020144

Mediano, P. A. M., Rosas, F. E., Farah, J. C., Shanahan, M., Bor, D., and Barrett, A. B. (2021). Integrated information as a common signature of dynamical and information-processing complexity. *arXiv:2106.10211 [q-bio.NC]*.

Montúfar, G., Ghazi-Zahedi, K., and Ay, N. (2015). A theory of cheap control in embodied systems. *PLoS Comput. Biol*. 11:e1004427. doi: 10.1371/journal.pcbi.1004427

Müller, V., and Hoffmann, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artif. Life* 23, 1–24. doi: 10.1162/ARTL_a_00219

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

Sporns, O., and Pegors, T. K. (2004). "Information-theoretical aspects of embodied artificial intelligence," in *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*, eds F. Iida, R. Pfeifer, L. Steels and Y. Kuniyoshi (Berlin; Heidelberg: Springer Berlin Heidelberg), 74–85.

Todorov, E., and Jordan, M. (2002). Optimal feedback control as a theory of motor coordination. *Nat. Neurosci*. 5, 1226–1235. doi: 10.1038/nn963

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

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

Tononi, G., and Edelman, G. M. (1998). Consciousness and complexity. *Science* 282, 1846–1851. doi: 10.1126/science.282.5395.1846

Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. U.S.A*. 91, 5033–5037. doi: 10.1073/pnas.91.11.5033

Touchette, H., and Lloyd, S. (2004). Information-theoretic approach to the study of control systems. *Physica A* 331, 140–172. doi: 10.1016/j.physa.2003.09.007

Toussaint, M. (2009). Probabilistic inference as a model of planned behavior. *Künstliche Intell*. 23, 23–29. Available online at: https://ipvs.informatik.uni-stuttgart.de/mlr/papers/09-toussaint-KI.pdf

Toussaint, M., Charlin, L., and Poupart, P. (2008). "Hierarchical pomdp controller optimization by likelihood maximization," in *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence* (Helsinki), 562–570.

Toussaint, M., Harmeling, S., and Storkey, A. (2006). *Probabilistic inference for solving (po)mdps*. Technical Report 934, School of Informatics, University of Edinburgh.

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

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

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

frontiers in Psychology

PERSPECTIVE
published: 23 December 2021
doi: 10.3389/fpsyg.2021.716976

Cognitive Models of Limb Embodiment in Structurally Varying Bodies: A Theoretical Perspective

Adna Bliek [1]*, *Robin Bekrater-Bodmann* [2] *and Philipp Beckerle* [1]

[1] *Chair of Autonomous Systems and Mechatronics, Department of Electrical Engineering, Faculty of Engineering, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany,* [2] *Department of Psychosomatic Medicine and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*

Using the seminal rubber hand illusion and related paradigms, the last two decades unveiled the multisensory mechanisms underlying the sense of limb embodiment, that is, the cognitive integration of an artificial limb into one's body representation. Since also individuals with amputations can be induced to embody an artificial limb by multimodal sensory stimulation, it can be assumed that the involved computational mechanisms are universal and independent of the perceiver's physical integrity. This is anything but trivial, since experimentally induced embodiment has been related to the embodiment of prostheses in limb amputees, representing a crucial rehabilitative goal with clinical implications. However, until now there is no unified theoretical framework to explain limb embodiment in structurally varying bodies. In the present work, we suggest extensions of the existing Bayesian models on limb embodiment in normally-limbed persons in order to apply them to the specific situation in limb amputees lacking the limb as physical effector. We propose that adjusted weighting of included parameters of a unified modeling framework, rather than qualitatively different model structures for normally-limbed and amputated individuals, is capable of explaining embodiment in structurally varying bodies. Differences in the spatial representation of the close environment (peripersonal space) and the limb (phantom limb awareness) as well as sensorimotor learning processes associated with limb loss and the use of prostheses might be crucial modulators for embodiment of artificial limbs in individuals with limb amputation. We will discuss implications of our extended Bayesian model for basic research and clinical contexts.

Keywords: bodily illusions, embodiment, structurally varying bodies, cognitive model, rubber limb illusion

OPEN ACCESS

Edited by:
Bernhard Hommel,
University Hospital Carl Gustav Carus,
Germany

Reviewed by:
Karthik Soman,
University of California, San Francisco,
United States
Shu Imaizumi,
Ochanomizu University, Japan

****Correspondence:***
Adna Bliek
adna.bliek@fau.de

Specialty section:
This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

Received: *29 May 2021*
Accepted: *01 December 2021*
Published: *23 December 2021*

Citation:
Bliek A, Bekrater-Bodmann R and Beckerle P (2021) Cognitive Models of Limb Embodiment in Structurally Varying Bodies: A Theoretical Perspective.
Front. Psychol. 12:716976.
doi: 10.3389/fpsyg.2021.716976

1. INTRODUCTION

Setups such as the rubber limb illusion (RLI) (e.g., Botvinick and Cohen, 1998; Flögel et al., 2016) and related paradigms (Riemer et al., 2019) have been comprehensively used to study the embodiment of artificial limbs in normally-limbed participants. In this context, "embodiment" refers to the cognitive integration of an external object into one's body representation (Longo et al., 2008; Makin et al., 2017). In the RLI, both a participant's real but hidden limb as well as a visible artificial counterpart are touched synchronously, inducing the perception that the artificial limb belongs to the participant's body. After successful RLI induction, participants tend to locate their hidden limb closer to the rubber limb than before, a phenomenon termed "proprioceptive drift" (e.g., Botvinick and Cohen, 1998; Longo et al., 2008), which has been interpreted as proprioceptive re-calibration of the body coordinates (Botvinick and Cohen, 1998). The vividness of the RLI has been found to depend on various parameters, such as the degree of synchrony between visual and tactile stimulation (Bekrater-Bodmann et al., 2014) or visual features such as color and shape of the artificial limb (Tsakiris et al., 2010; Farmer et al., 2012). Another important factor for eliciting the RLI is whether the artificial limb is placed in the individual's limb-centered peripersonal space (PPS) (Lloyd, 2007), i.e., the intermediate surroundings of a limb, within the limits of which the integration of multimodal stimuli is facilitated (Serino, 2019). For humans, limb-centered PPS boundaries of about 30 cm for the upper limb (Lloyd, 2007) and 70 cm for the lower limb (Stone et al., 2018) have been reported.

Traditionally, the embodiment experiences elicited in the RLI have been assumed to rely exclusively on bottom-up processes (Botvinick and Cohen, 1998; Armel and Ramachandran, 2003), emphasizing a three-way interaction between vision, touch, and proprioception, which—in a connectionist tradition—leads to perceived merging of tactile and visual inputs by distortions of the position sense (Armel and Ramachandran, 2003). However, this purely bottom-up perspective is not compatible with the growing number of empirical evidence on the principles underlying limb embodiment (Litwin, 2020), as earlier described top-down modulating factors, e.g., the PPS, have been shown to influence the vividness of the RLI.

In the light of recent advances in Bayesian modeling of sensory integration (Körding and Wolpert, 2006; Körding et al., 2007; Berniker and Kording, 2011), the processes involved in the RLI have been proposed to be better modeled as multisensory combination based on probabilistic principles (Samad et al., 2015; Schürmann et al., 2019b; Litwin, 2020; Shams and Beierholm, 2021). In this view, embodiment of an external object takes place when multimodal sensory inputs are (falsely) interpreted as being caused by the same external event. Bayesian modeling has first been used by Samad et al. (2015) to predict the strength of the RLI for the hand explaining the induction of embodiment in a traditional bottom-up fashion. In their model, the induction of embodiment depends on whether a cognitive system infers common or independent causes of visual and somatosensory signals, resulting in a re-calibration of proprioceptive coordinates. The combination of sensory signals would then depend on the relative probabilities of these two posterior hypotheses derived from their prior probabilities and likelihood of sensory signals. By empirical testing of hypotheses deduced from their model, Samad et al. (2015) found strong evidence for Bayesian probabilistic processing underlying the embodiment of an artificial limb. However, a recent article by Schubert and Endres (2021) highlighted flaws in the unrealistic wide choice of the prior distributions of the model. They could not recreate realistic results using their improved prior distributions given the current model structure. Additionally, Schürmann et al. (2019b) showed that informed priors outperform the originally used uniform priors.

Crucially, the RLI can also be induced in individuals with limb amputations (e.g., Ehrsson et al., 2008) which is why this setup has been proposed to be a model for certain processes involved in the embodiment of prostheses as well. Successful embodiment of a prosthesis is important as the amputation of a limb severely disrupts a person's physical integrity. There are preliminary reports that most individuals with amputations can achieve embodiment of their prosthesis (Bekrater-Bodmann, 2020) the processes of which have been related to positive clinical outcomes (e.g., Imaizumi et al., 2016; Bekrater-Bodmann, 2021; Bekrater-Bodmann et al., 2021). The psychometric structure behind experimentally-induced short-termed RLI experiences in normally-limbed participants (Longo et al., 2008) and real-life long-termed prosthesis embodiment in limb amputees (Bekrater-Bodmann, 2020) show substantial qualitative similarity, which is remarkable, given the striking differences in the participant's physical integrity. Although there is reason to assume that differential neurocognitive mechanisms contribute to embodiment experiences in the RLI and prosthesis use, with the latter probably relying on long-term sensorimotor learning rather than short-term multimodal sensory combination (cf., Zbinden and Ortiz-Catalan, 2021), the psychometric similarities suggest at least partly overlapping, potentially Bayesian processes.

However, the question arises how bodily self-experiences in general and prosthesis embodiment in particular can be theoretically explained in a unified fashion taking into account and improving on the currently used Bayesian modeling approaches. A unified modeling framework could be a step toward prediction of factors improving embodiment of artificial limbs and could thus improve user experience. The authors of the present article propose a 2-fold extension of the current modeling approaches in accordance with the upper two levels of cognitive modeling proposed by Marr (1982), which has been proposed in earlier research to describe the different underlying task of modeling approaches (e.g., Schürmann and Beckerle, 2020; Shams and Beierholm, 2021). Firstly, starting on the computational theory level, we propose to improve the current model structure, and extend the models for structurally varying bodies taking into account individual differences in perception of embodiment. Secondly, on the algorithmic level, we propose to incorporate top-down modulating factors in the priors of the cognitive models, according to Litwin (2020).

2. LIMB EMBODIMENT IN STRUCTURALLY VARYING BODIES

Normally-limbed and amputated individuals differ in important representational and perceptual characteristics, which have to be considered when cognitive modeling is applied to the processes underlying artificial limb embodiment. Thus, limb amputees often report the presence of a phantom limb (Kooijman et al., 2000), i.e., the persistent perception of a body part that has been removed. The proprioceptive presence of a phantom limb can be made use of in the induction of embodiment: in some individuals with amputations, tactile stimulation applied to the residual limb can trigger a touch sensation in the phantom limb, known as "referred sensations", which might be a consequence of neuroplastic changes in the somatotopic body maps in the brain (Ramachandran et al., 1992). If the location of the elicited sensations in the phantom corresponds to the visual location of touch applied to the artificial limb, embodiment experiences can be facilitated (Ehrsson et al., 2008). Furthermore, there is preliminary evidence that prostheses interact with the phantom limb in terms of perceptual co-location (the phantom "occupies" the space of the prosthesis; Giummarra et al., 2008) which might also foster the embodiment of the prosthetic device. Postural phantom limb disturbances, however, could interfere with the incorporation of the artificial limb and consequently reduce embodiment (cf., Foell et al., 2014).

Moreover, limb amputation is associated with a shrinkage in the extent of PPS representation, with a shift of its boundaries toward the stump (Canzoneri et al., 2013), which might explain why prosthesis embodiment is strong for long residual limbs and low for short ones (Bekrater-Bodmann, 2020): in short residual limbs, the prosthesis might "stick out" of the PPS boundaries which interferes with its embodiment (cf., Lloyd, 2007). Whether or not phantom limb perceptions are associated with normal PPS extent, however, remains unknown.

3. MODELS PREDICTING EMBODIMENT FOR STRUCTURALLY VARYING BODIES

Given both the perceptual similarities and potential mechanistic differences, i.e., integration of multimodal sensory input vs. sensorimotor learning processes, between short-term and long-term embodiment in normally-limbed and amputated individuals, the combination of Bayesian and connectionist models and the modulation of priors seem promising for the prediction of experiences in both groups. However, it is currently unclear how these models should be combined or adapted. Current embodiment models do not cover structural body varieties, e.g., limb presence or absence, since priors do not take into account inter-individual differences in body representation, e.g., differences in PPS extent and different underlying mechanisms. One crucial issue could relate to quantitatively different weighting of certain sensorimotor factors in normally-limbed vs. amputated individuals, while the structure of the model itself remains unaffected. This might allow for the integration of different PPS representations in amputated and normally-limbed bodies, as preliminary indicated by Canzoneri et al. (2013). Samad et al. (2015) highlighted that their proposed framework is extendable to incorporate such additional variables by adding individual prior knowledge, e.g., by adding tactile, proprioceptive, and visual priors and adapting their sensitivity to the individual or a group of people. The importance of prior knowledge is highlighted by recent evidence suggesting that the prediction quality for behavioral correlates of the RLI can be enhanced by entering informed priors to the probabilistic model (Schürmann et al., 2019b). Litwin (2020) further opts for the inclusion of individual dispersions of coupling priors for modeling the potentially important, but largely neglected, top-down effects. Thus, beyond bottom-up processes, the human cognitive system seems to use prior knowledge of cross-modal correlations, e.g., the correlation between visual and tactile stimulation, to modulate sensory integration in PPS (Parise et al., 2012, 2013), which might be subject to individual sensorimotor experiences and learning. The embodiment of an artificial limb has been linked to remapping of PPS boundaries to the location of the artificial limb (Brozzoli et al., 2012), suggesting that its representation is shaped by top-down influences.

To factor in individual representational differences, we suggest to improve the existing Bayesian models with respect to estimation accuracy, specifications for individual users, and online capabilities. **Figure 1** provides a conceptual perspective of an extended framework to include individual differences. Starting from multisensory models aiming to predict the perceived limb location, i.e., the proprioceptive drift (Samad et al., 2015; Schürmann et al., 2019b), extending the models with sensorimotor information in order to cover more complex behavioral and psychological outcomes. To this end, we suggest extending the current Bayesian model considering the upper two levels of Marr (1982): the computational theory level, describing what a system is doing and what functions are needed to complete this goal, and the algorithmic level, outlining how the system could be implemented (Marr, 1982; Dennett, 1987).

The goal of the proposed framework is to estimate the embodiment of an artificial limb for an individual taking into account structural differences of their bodies. We suggest that this goal can be realized by combining established models of multisensory integration with models of perception and higher cognition, and extending the overall framework by experience-modulated priors. These changes are indicated in the addition of the model of cognition, the models of sensation and perception, and the top-down modulation in **Figure 1**. The added priors would not be individualized but represent general influences of experience, i.e., irrespective of structural body variations. Computationally, this could be covered by a top-down modulation to predict influences of previous experiences on prior couplings, e.g., visuo-tactile integration or sensorimotor learning, using the implementation of learning-based models of inter- and intramodal sensory signals (Van Dam et al., 2014; Parise, 2016; Noel et al., 2018; Litwin, 2020; Press et al., 2020).

On the algorithmic level, we propose to extend the approaches and mechanisms in the submodels of sensation and perception, cognition, and top-down modulation, see algorithmic approaches in **Figure 1**. The model of cognition adds psychometric measures of embodiment, e.g., perceived agency and body ownership, in order to include individual perceptual outcomes in addition to the proprioceptive measures. To include more individualized information in the model of sensation and perception, Bayesian and connectionist methods as well as predictive coding are promising for the perceptual submodels, e.g., by adding sensorimotor learning, (Thomas and McClelland, 2008; Clark, 2013; Samad et al., 2015; Schürmann et al., 2019a,b). The top-down connection between the model of cognition, and the model of sensation and perception is adding experience-modulated priors (cf., Ingram et al., 2017), incorporating recent evidence for top-down modulation of adaptive sensory representations in the brain (Makino et al., 2016). We propose adding a top-down modulation of priors to incorporate information about individual PPS, visuo-tactile

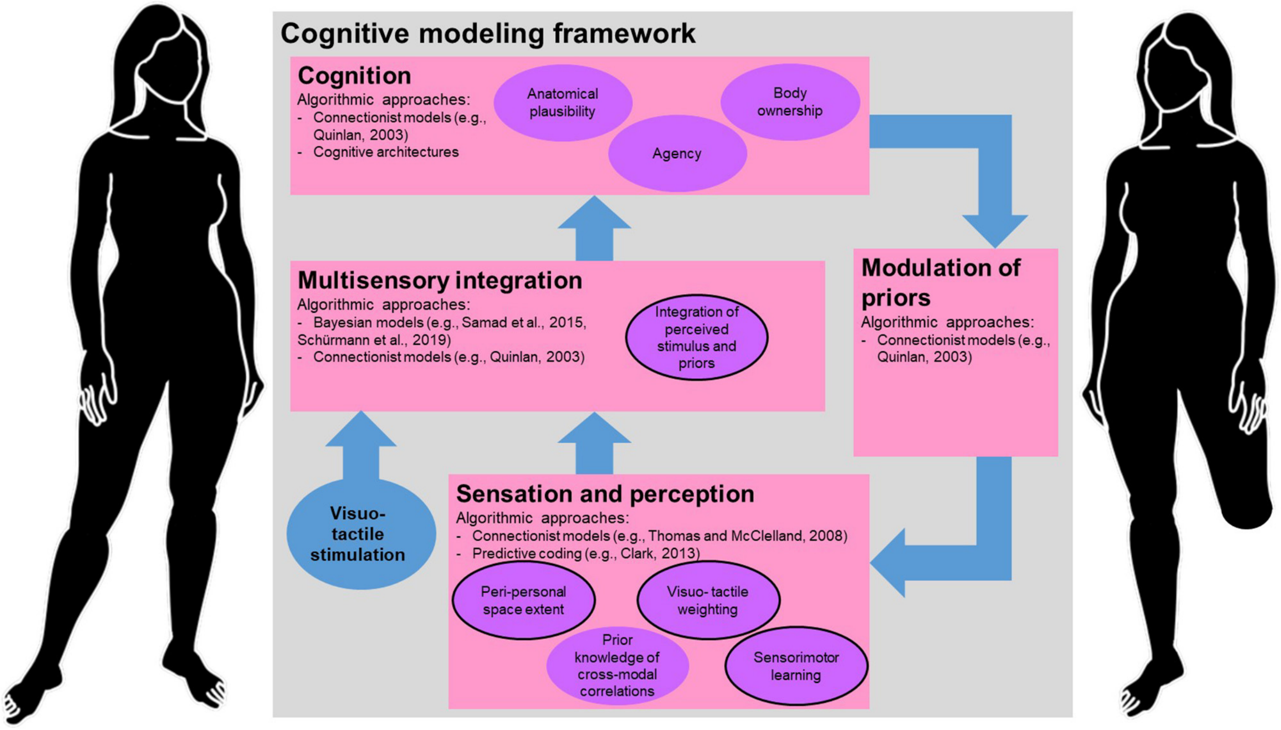

Cognitive modeling framework
Cognition
Algorithmic approaches:
- Connectionist models (e.g., Quinlan, 2003)
- Cognitive architectures
Anatomical plausibility
Body ownership
Agency
Multisensory integration
Algorithmic approaches:
- Bayesian models (e.g., Samad et al., 2015, Schürmann et al., 2019)
- Connectionist models (e.g., Quinlan, 2003)
Integration of perceived stimulus and priors
Modulation of priors
Algorithmic approaches:
- Connectionist models (e.g., Quinlan, 2003)
Visuo-tactile stimulation
Sensation and perception
Algorithmic approaches:
- Connectionist models (e.g., Thomas and McClelland, 2008)
- Predictive coding (e.g., Clark, 2013)
Peri-personal space extent
Visuo-tactile weighting
Prior knowledge of cross-modal correlations
Sensorimotor learning

FIGURE 1 | The proposed cognitive modeling framework suggests a conceptual structure representing embodiment mechanisms on the computational level. Submodels (pink boxes) are connected by information flows (blue arrows). The submodels contain suggestions for algorithmic approaches to realize the functions represented in the purple ellipses. Among those, mechanisms that appear important for long-term differences in embodiment, i.e., sensorimotor learning in individuals with amputations, are outlined. The framework further suggests how to model the integration of visuo-tactile with proprioceptive information. Bottom-up information is used as input for the model of cognition from the perceived stimulation as well as from prior knowledge. The prior knowledge is modulated using top-down information from earlier experiences.

weighting, sensorimotor learning and prior knowledge of cross-modal correlations, indicated by the ellipses in the submodel of sensation and perception in **Figure 1**. This pathway could be realized in a connectionist fashion, e.g., by the implementation of artificial neural networks (Quinlan, 2003; Zhong, 2015). Artificial neural networks, as well as network architecture in the brain (Graham, 1982), use feedback information to update the weights of the connections between neurons. This process makes them adaptable to individual differences, while also modeling processes that are valid on group level. These approaches appear to be particularly promising for limb amputees who are characterized by high variability in sensorimotor experiences related to the use of prostheses.

To ensure accurate models for structurally varying bodies, the suggested algorithmic model adaptations should be performed iteratively using human-in-the-loop experiments with individuals with structurally varying bodies, e.g., people with/without amputation, to verify and adapt the implemented models and priors. We postulate that the overall cognitive modeling framework should be generally applicable to structurally varying bodies at computational level. The methods selected on algorithmic level might be identical, but should vary in the parameterization that represents individual effects, e.g., artificial neural network weights.

4. CONCLUSION

Both the similarities and the differences of limb embodiment in individuals with structurally varying bodies show a need for an extension of currently used cognitive models for normally-limbed people. These models should be adapted to consider individual limb differences by incorporating further parameters such as the peripersonal space and adapting the weighting of included parameters iteratively to the individual. Such extensions could not only help to explain and predict embodiment of prostheses but also highlight individual factors that facilitate or hinder embodiment of rehabilitative devices in general.

The current research points toward prior sensorimotor experiences and the peripersonal space extent taking influence on the embodiment of (artificial) limbs. Thus, we advocate to create a cognitive modeling framework that extends current approaches with top-down modulations to represent individual structural and other representational differences and make algorithmic suggestions to realize its implementation, e.g., using artificial neural networks or cognitive architectures.

Furthermore, modeling embodiment for both individuals with and without amputation will enable the characterization of the variability (or invariability) of different parameters of the model, e.g., the sensitivity of priors or the importance of used prior knowledge in cognitive architectures or artificial neural networks. In other words, the comparison of the models' dynamics for structurally varying bodies will reveal to which degree the bodily self is subject to plastic adaptions in response to structural alterations of the physical body. To accurately model the variability in the processes involved in limb embodiment, experiments with participants with and without amputations will be needed before adapting the models to inform theoretical considerations. Supported by neuropsychological research, the proposed modeling approaches might foster our understanding of the mechanisms underlying limb embodiment and the predictive power of cognitive models, which might in turn be used to improve the design and control of assistive devices.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG; BE 5723/4-1) and the Volkswagen-Foundation (Az. 9B 007).

REFERENCES

Armel, K. C., and Ramachandran, V. S. (2003). Projecting sensations to external objects: evidence from skin conductance response. *Proc. R. Soc. B Biol. Sci*. 270, 1499–1506. doi: 10.1098/rspb.2003.2364

Bekrater-Bodmann, R. (2020). Perceptual correlates of successful body–prosthesis interaction in lower limb amputees: psychometric characterisation and development of the prosthesis embodiment scale. *Sci. Rep*. 10, 1–13. doi: 10.1038/s41598-020-70828-y

Bekrater-Bodmann, R. (2021). Factors associated with prosthesis embodiment and its importance for prosthetic satisfaction in lower limb amputees. *Front. Neurorobot*. 14:604376. doi: 10.3389/fnbot.2020.604376

Bekrater-Bodmann, R., Foell, J., Diers, M., Kamping, S., Rance, M., Kirsch, P., et al. (2014). The importance of synchrony and temporal order of visual and tactile input for illusory limb ownership experiences–an fmri study applying virtual reality. *PLoS ONE* 9:e87013. doi: 10.1371/journal.pone.0087013

Bekrater-Bodmann, R., Reinhard, I., Diers, M., Fuchs, X., and Flor, H. (2021). Relationship of prosthesis ownership and phantom limb pain: results of a survey in 2383 limb amputees. *Pain* 162, 630–640. doi: 10.1097/j.pain.0000000000002063

Berniker, M., and Kording, K. (2011). Bayesian approaches to sensory integration for motor control. *Wiley Interdiscipl. Rev. Cogn. Sci.* 2, 419–428. doi: 10.1002/wcs.125

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel'touch that eyes see. *Nature* 391, 756–756. doi: 10.1038/35784

Brozzoli, C., Gentile, G., and Ehrsson, H. H. (2012). That's near my hand! parietal and premotor coding of hand-centered space contributes to localization and self-attribution of the hand. *J. Neurosci.* 32, 14573–14582. doi: 10.1523/JNEUROSCI.2660-12.2012

Canzoneri, E., Marzolla, M., Amoresano, A., Verni, G., and Serino, A. (2013). Amputation and prosthesis implantation shape body and peripersonal space representations. *Sci. Rep.* 3:2844. doi: 10.1038/srep02844

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

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Ehrsson, H. H., Rosén, B., Stockselius, A., Ragnö, C., Köhler, P., and Lundborg, G. (2008). Upper limb amputees can be induced to experience a rubber hand as their own. *Brain* 131, 3443–3452. doi: 10.1093/brain/awn297

Farmer, H., Tajadura-Jiménez, A., and Tsakiris, M. (2012). Beyond the colour of my skin: how skin colour affects the sense of body-ownership. *Conscious. Cogn.* 21, 1242–1256. doi: 10.1016/j.concog.2012.04.011

Flögel, M., Kalveram, K. T., Christ, O., and Vogt, J. (2016). Application of the rubber hand illusion paradigm: comparison between upper and lower limbs. *Psychol. Res.* 80, 298–306. doi: 10.1007/s00426-015-0650-4

Foell, J., Bekrater-Bodmann, R., Diers, M., and Flor, H. (2014). Mirror therapy for phantom limb pain: brain changes and the role of body representation. *Eur. J. Pain* 18, 729–739. doi: 10.1002/j.1532-2149.2013.00433.x

Giummarra, M. J., Gibson, S. J., Georgiou-Karistianis, N., and Bradshaw, J. L. (2008). Mechanisms underlying embodiment, disembodiment and loss of embodiment. *Neurosci. Biobehav. Rev.* 32, 143–160. doi: 10.1016/j.neubiorev.2007.07.001

Graham, J. (1982). Some topographical connections of the striate cortex with subcortical structures in macaca fascicularis. *Exp. Brain Res.* 47, 1–14. doi: 10.1007/BF00235880

Imaizumi, S., Asai, T., and Koyama, S. (2016). Embodied prosthetic arm stabilizes body posture, while unembodied one perturbs it. *Conscious. Cogn.* 45, 75–88. doi: 10.1016/j.concog.2016.08.019

Ingram, J. N., Sadeghi, M., Flanagan, J. R., and Wolpert, D. M. (2017). An error-tuned model for sensorimotor learning. *PLoS Comput. Biol.* 13:e1005883. doi: 10.1371/journal.pcbi.1005883

Kooijman, C. M., Dijkstra, P. U., Geertzen, J. H., Elzinga, A., and Van der Schans, C. P. (2000). Phantom pain and phantom sensations in upper limb amputees: an epidemiological study. *Pain* 87, 33–41. doi: 10.1016/S0304-3959(00)00264-5

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE* 2:e943. doi: 10.1371/journal.pone.0000943

Körding, K. P., and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends Cogn. Sci.* 10, 319–326. doi: 10.1016/j.tics.2006.05.003

Litwin, P. (2020). Extending Bayesian models of the rubber hand illusion. *Multisensory Res.* 33, 127–160. doi: 10.1163/22134808-20191440

Lloyd, D. M. (2007). Spatial limits on referred touch to an alien limb may reflect boundaries of visuo-tactile peripersonal space surrounding the hand. *Brain Cogn.* 64, 104–109. doi: 10.1016/j.bandc.2006.09.013

Longo, M. R., Schüür, F., Kammers, M. P., Tsakiris, M., and Haggard, P. (2008). What is embodiment? A psychometric approach. *Cognition* 107, 978–998. doi: 10.1016/j.cognition.2007.12.004

Makin, T. R., de Vignemont, F., and Faisal, A. A. (2017). Neurocognitive barriers to the embodiment of technology. *Nat. Biomed. Eng.* 1, 1–3. doi: 10.1038/s41551-016-0014

Makino, H., Hwang, E. J., Hedrick, N. G., and Komiyama, T. (2016). Circuit mechanisms of sensorimotor learning. *Neuron* 92, 705–721. doi: 10.1016/j.neuron.2016.10.029

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge: MIT Press.

Noel, J.-P., Samad, M., Doxon, A., Clark, J., Keller, S., and Di Luca, M. (2018). Peripersonal space as a prior in coupling visual and proprioceptive signals. *Sci. Rep.* 8, 1–15. doi: 10.1038/s41598-018-33961-3

Parise, C. V. (2016). Crossmodal correspondences: standing issues and experimental guidelines. *Multisensory Res.* 29, 7–28. doi: 10.1163/22134808-00002502

Parise, C. V., Harrar, V., Ernst, M. O., and Spence, C. (2013). Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisensory Res.* 26, 307–316. doi: 10.1163/22134808-00002417

Parise, C. V., Spence, C., and Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Curr. Biol.* 22, 46–49. doi: 10.1016/j.cub.2011.11.039

Press, C., Kok, P., and Yon, D. (2020). The perceptual prediction paradox. *Trends Cogn. Sci.* 24, 13–24. doi: 10.1016/j.tics.2019.11.003

Quinlan, P. T. (2003). *Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks*. Taylor & Francis.

Ramachandran, V. S., Rogers-Ramachandran, D., Stewart, M., and Pons, T. P. (1992). Perceptual correlates of massive cortical reorganization. *Science* 258, 1159–1159. doi: 10.1126/science.1439826

Riemer, M., Trojan, J., Beauchamp, M., and Fuchs, X. (2019). The rubber hand universe: on the impact of methodological differences in the rubber hand illusion. *Neurosci. Biobehav. Rev.* 104, 268–280. doi: 10.1016/j.neubiorev.2019.07.008

Samad, M., Chung, A. J., and Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLoS ONE* 10:e0117178. doi: 10.1371/journal.pone.0117178

Schubert, M., and Endres, D. (2021). More plausible models of body ownership could benefit virtual reality applications. *Computers* 10:108. doi: 10.3390/computers10090108

Schürmann, T., and Beckerle, P. (2020). Personalizing human-agent interaction through cognitive models. *Front. Psychol.* 11:561510. doi: 10.3389/fpsyg.2020.561510

Schürmann, T., Mohler, B. J., Peters, J., and Beckerle, P. (2019a). How cognitive models of human body experience might push robotics. *Front. Neurorobot.* 13:14. doi: 10.3389/fnbot.2019.00014

Schürmann, T., Vogt, J., Christ, O., and Beckerle, P. (2019b). The Bayesian causal inference model benefits from an informed prior to predict proprioceptive drift in the rubber foot illusion. *Cogn. Process.* 20, 447–457. doi: 10.1007/s10339-019-00928-9

Serino, A. (2019). Peripersonal space (pps) as a multisensory interface between the individual and the environment, defining the space of the self. *Neurosci. Biobehav. Rev.* 99, 138–159. doi: 10.1016/j.neubiorev.2019.01.016

Shams, L., and Beierholm, U. (2021). Bayesian causal inference: a unifying neuroscience theory. *PsyArXiv [Preprint]*. doi: 10.31234/osf.io/xpz6n

Stone, K. D., Kandula, M., Keizer, A., and Dijkerman, H. C. (2018). Peripersonal space boundaries around the lower limbs. *Exp. Brain Res.* 236, 161–173. doi: 10.1007/s00221-017-5115-0

Thomas, M. S. C., and McClelland, J. L. (2008). *Connectionist Models of Cognition*. New York, NY: Cambridge Handbooks in Psychology; Cambridge University Press, 23–58.

Tsakiris, M., Carpenter, L., James, D., and Fotopoulou, A. (2010). Hands only illusion: multisensory integration elicits sense of ownership for body parts but not for non-corporeal objects. *Exp. Brain Res.* 204, 343–352. doi: 10.1007/s00221-009-2039-3

Van Dam, L., Parise, C., and Ernst, M. (2014). "Modeling multisensory integration," in *Sensory Integration and the Unity of Consciousness*, eds D. J. Bennett and C. S. Hill (Cambridge, MA: MIT Press), 209–230.

Zbinden, J., and Ortiz-Catalan, M. (2021). The rubber hand illusion is a fallible method to study ownership of prosthetic limbs. *Sci. Rep.* 11, 1–11. doi: 10.1038/s41598-021-83789-7

Zhong, J. (2015). *Artificial neural models for feedback pathways for sensorimotor integration* (Ph.D thesis). Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, Hamburg, Germany.

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

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

REVIEW
published: 18 March 2022
doi: 10.3389/fpsyg.2022.714464

Perspective Taking and Avatar-Self Merging

Jochen Müsseler, Sophia von Salm-Hoogstraeten and Christian Böffel*

Institute of Psychology, Work and Engineering Psychology, RWTH Aachen University, Aachen, Germany

Today, avatars often represent users in digital worlds such as in video games or workplace applications. Avatars embody the user and perform their actions in these artificial environments. As a result, users sometimes develop the feeling that their self merges with their avatar. The user realizes that they are the avatar, but the avatar is also the user—meaning that avatar's appearance, character, and actions also affect their self. In the present paper, we first introduce the event-coding approach of the self and then argue based on the reviewed literature on human-avatar interaction that a self-controlled avatar can lead to avatar-self merging: the user sets their own goals in the virtual environment, plans and executes the avatar's actions, and compares the predicted with the actual motion outcomes of the avatar. This makes the user feel body ownership and agency over the avatar's action. Following the event-coding account, avatar-self merging should not be seen as an all-or-nothing process, but rather as a continuous process to which various factors contribute, including successfully taking the perspective of the avatar. Against this background, we discuss affective, cognitive, and visuo-spatial perspective taking of the avatar. As evidence for avatar-self merging, we present findings showing that when users take the avatar's perspective, they can show spontaneous behavioral tendencies that run counter to their own.

Keywords: perspective taking, minimal self, avatar-self merging, Theory of Event Coding, avatar embodiment, spatial compatibility, ownership, agency

Edited by:
Bernhard Hommel,
University Hospital Carl Gustav Carus, Germany

Reviewed by:
Dennis Küster,
University of Bremen, Germany
Stephan Alexander Verschoor,
Leiden University, Netherlands

****Correspondence:***
Jochen Müsseler
muesseler@psych.rwth-aachen.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *25 May 2021*
Accepted: *18 February 2022*
Published: *18 March 2022*

Citation:
Müsseler J, von Salm-Hoogstraeten S and Böffel C (2022) Perspective Taking and Avatar-Self Merging. Front. Psychol. 13:714464. doi: 10.3389/fpsyg.2022.714464

INTRODUCTION

Originally, the term avatar referred to a deity of Indian mythology who descended to earth in a human appearance with the aim to enable mankind new insights, self-discoveries, and self-realizations. Nowadays, this term is transferred to virtual environments with abstract 2D outlines of individuals (e.g., a gravatar, Wolf and Henley, 2017) and 3D animated artificial characters (e.g., as illustrated in the movie "Avatar" by James Cameron, 2009). They are understood to either represent a completely independent artificial character or to act in place of a user in a virtual environment (Pan and Hamilton, 2018). In the present context, we refer to the term avatar in the latter sense. An avatar is understood as a (social) tool, as an extended "arm" of the user in video games and—increasingly also—in workplace applications. It enables the user to realize own intentions and goals in the virtual environment.

After intensive training and engagement with such an avatar, after navigating and interacting with it in the virtual environment, some users develop the feeling that they are integrating the avatar into their selves. They may even get the feeling of becoming one with it—a process we refer to as avatar-self merging (Böffel, 2021).[1] We prefer this term as it captures the interactive influences of avatar and user. In gaming and virtual reality, the user realizes that they are the avatar, but the avatar is also them—meaning that its appearance, character, and actions affect also their self (Böffel, 2021). The avatar also does not replace only body parts, as various body-ownership illusions (e.g., Kilteni et al., 2015) and some prosthetic studies (e.g., Bekrater-Bodmann, 2020) suggest. For instance, arm amputees often report that their tool, the prosthetic arm, becomes a part of themselves after a period of training. We will argue that avatar-self merging goes beyond this because it emphasizes the interactive social component between user and avatar that pure ownership of body parts lacks.

In this paper, we synthesize existing studies and theories surrounding the user-avatar interaction and argue that controlling an avatar and taking its perspective is best described by the concept of avatar-self merging. We examine the conditions that facilitate but also constrain avatar-self merging. Before we do that, we need to clarify what the self is about and consider a prerequisite of successful avatar-self merging, namely, to successfully take the perspective of the (virtual) character.

THE ENRICHMENT OF THE SELF THROUGH AN AVATAR: AVATAR-SELF MERGING

Scientifically, two components are associated with the concept of the self: the minimal self and the narrative self (e.g., Gallagher, 2000). The minimal self is considered as the experience of our self in the here and now. Like other authors (e.g., Gallese and Sinigaglia, 2010; Hommel, 2018, 2021), we understand it as action-oriented, in the sense that it arises from our sensorimotor interactions with the environment. In contrast, the narrative self reflects our life experiences, which—among other events—contribute to our personal identity. It is assumed to need memory and language to be established.

Since the present context is primarily concerned with the sensorimotor interactions of users and their avatars in the virtual environment, we focus on the minimal self. More specifically, the interactions are assumed to give rise to the experiences of perceived body ownership and perceived agency, which in turn are seen as the constituting elements of the minimal self (see also Verschoor and Hommel, 2017). Perceived body ownership is understood as a person's impression that their body belongs to them and is distinct from their environment. Healthy persons usually feel their hand belongs to them, but they may also perceive a rubber hand in front of them as part of their body if that rubber hand is oriented like their hand and stroked simultaneously with it (so-called rubber-hand illusion, Botvinick and Cohen, 1998; Costantini and Haggard, 2007).

Perceived agency refers to the impression of being the originator of an action and of controlling events in the environment with this action. This impression of being an agent arises when we lift a beverage with our hand, for example, but also when this is done indirectly with a mechanic gripping tool. In the latter case, the cognitive and motor performances (force, movement distance, etc.) can be completely different; nevertheless, we attribute the lifting action to us (e.g., Sutter et al., 2013).

Perceived ownership and perceived agency are seen to be intimately linked, modulated by each other (de Haan and de Bruin, 2010), and influenced by the same manipulations (Ma et al., 2019, 2021). Thus, it is not completely clear what separate contributions both concepts make to the minimal self. A further problem is that they are often gathered with subjective questionnaires, which are known to be prone to errors and biases. This has led to the concept of the minimal self being burdened with a certain degree of fuzziness.

Last but not least, there was a lack of ideas about how to conceive the representation of the self in the cognitive system. In this regard, Hommel (2018, see also Hommel, 2021) developed a promising approach in recent papers. He started from the Theory of Event Coding (TEC, Hommel et al., 2001) and assumes that the representation of the self and the representation of the others are event files consisting of a bundle of feature codes at a given moment (color, shape, location, but also motor properties and goals, etc.). In principle, the representation of the self (the minimal self) and the representation of the others do not differ, but the self has (1) preferential and, in part, exclusive access to our sensations (e.g., with regard to proprioceptive sensations). (2) The ideomotor principle as an integral part of TEC enables the planning and execution of motor activities and (3) the comparison between the predicted and actual motor outcomes allows us to judge fairly reliably whether we are the originator of an action or not. This lets us distinguish ourselves from the self of others.

Still, the event files of ourselves may also share features with the event files of others. A high degree of self-other overlap may promote mutual empathy, for instance (cf. Quintard et al., 2021). In the present context, such feature overlap is especially interesting when the other is an avatar. An increased self-avatar overlap is likely as the user sets the goals in the virtual environment, controls the avatar's actions, and compares the predicted with the actual motion outcomes of the avatar. This makes the user feel as if she is the originator of avatar's action, which might also lead to perceived body ownership. These are exactly the conditions that promote avatar-self merging.

The extent of self-avatar overlap is not fixed but varies with the user's traits and features and with the avatar's characteristics and action options. A user's personality (Dunn and Guadagno, 2012) or gender and race (Dunn and Guadagno, 2019), for example, predict which avatar they choose. In turn, the appearance of the avatar influences the user's behavior, and identification

[1] Other authors (e.g., Fribourg et al., 2020; Peck and Gonzalez-Franco, 2021) describe this feeling as avatar embodiment. However, in cognitive psychology, the term embodiment is used completely independent from artificial characters, instead it refers to the (theoretical) view that considers body states and actions as important or obligatory components of cognitive processes (e.g., Wilson, 2002). This was another reason not to use the term embodiment in the present context.

with the avatar increases with perceived interactivity (Hefner et al., 2007). Accordingly, and in contrast to other approaches, avatar-self merging describes a bi-directional process in which user and avatar influence each other. Furthermore, avatar-self merging is not seen as an all-or-nothing process but forms a continuum of varying intensities. Just as the extent of self-merging might be different between a plumber with their pliers and an arm amputee with their prosthesis, the difference is finally only gradual. Their tools, the pliers, and the prosthesis have become an integral part of their user's lives, make their intentions and goals achievable, expand their action space, and make impossible actions possible. An avatar similarly increases the user's action space and possibilities, but beyond that an avatar can be seen as a human(-like) being with its own appearance and character.

Successful avatar-self merging requires that the user puts themself in the situation of this character, that is, the user has to take its perspective. Perspective taking (PT) is an important process, when interacting with others. In its broader sense, it describes the ability to put oneself in the place of another person and to infer their mental states (e.g., percepts, feelings, beliefs, needs, and goals; Flavell et al., 1981; Steins and Wicklund, 1993; Birch et al., 2017). PT covers three mental aspects at least: affective PT (understanding another's emotions and affects, i.e., compassion or empathy), cognitive PT (understanding [unobservable] processes within a person, e.g., this person is lying), and visual-spatial PT (considering the visual–spatial perspective of another person; cf. Steins and Wicklund, 1993).[2] In the following, we discuss avatar-self merging against the background of affective, cognitive, and visual-spatial PT.

AFFECTIVE AND COGNITIVE PERSPECTIVE TAKING: ADOPTING THE AVATAR'S ASSIGNED CHARACTER

At first glance, acting with a self-controlled avatar in a virtual environment resembles a (social) situation in which a human observer attempts to infer the mental states of another person (here the avatar) in order to understand and predict its behavior. At second glance, as the avatar represents the user, the mental states of the avatar should be directly accessible to them—however, this does not mean that the assigned appearance and character of the avatar do not affect perspective taking.

Avatars are presented abstractly up to human-like. In some studies, avatars were found to be subjectively preferred, the more realistic they are (e.g., Fribourg et al., 2020). A more realistic avatar also seems to increase perceived body ownership (e.g., Latoschik et al., 2017), although this may not always be beneficial. Lugrin et al. (2015) reported that users feel stronger with a non-realistic but tough-looking avatar—a finding that is reflected in the so-called Proteus effect: Users adjust their behavior according to a randomly assigned appearance and/or character of an avatar. Yee and Bailenson (2007) showed that participants behaved in correspondence with stereotypes caused by the perception of their own avatar, for example, by being more confident when their avatar was taller. Similar effects have been demonstrated across different contexts, such as aggressive behavior (Ash, 2016), exercise habits (Fox and Bailenson, 2009), pro- and antisocial behavior (Yoon and Vargas, 2014), financial decisions (Hershfield et al., 2011), avatar's age (Beaudoin et al., 2020; Reinhard et al., 2020), and many more (for an overview see Ratan et al., 2020). There is also evidence that users adapt not only their behavior but also their mental attitudes to the avatar (Banakou et al., 2013).

Current explanations of the Proteus effect do not refer to self-merging. For example, Peña et al. (2009) attributed the Proteus effect to priming and inhibition processes triggered by the appearance of the avatar. Their assumption is that an aggressive-looking avatar primes an aggressive model and inhibits the inconsistent non-aggressive one and that without assuming a recourse to self-merging processes. However, explanations like priming and inhibition on the one hand and self-merging on the other are not mutually exclusive. Priming and inhibition refer to the processes, while self-merging refers to whether and to what extent the user feels that the avatar belongs to them or not. Thus, avatar-self merging may be indicated, when the user adapts their behavior to the appearance and character of an avatar.

VISUAL-SPATIAL PERSPECTIVE TAKING

The dominant sense of humans is vision, and so it is not surprising that PT also covers the ability to see the space around another person from its perspective. This visual-spatial perspective taking (VSPT) accounts for what the other person (here the avatar) sees and how they see it (Flavell, 1977), for instance, whether objects are (partially) occluded from their view or whether they can see something that the observer (here the user) is unable to see. Research on VSPT has its origin in developmental psychology. Flavell et al. (1981) distinguished between two developmental levels of VSPT. While at the earlier "level 1 VSPT," the child has insights into what objects are visible or occluded from the other's point of view, "level 2 VSPT" adds further insights how others perceive the world, including deviating distances and deviating relative positioning from one's own perspective (**Figure 1**). Level 2 VSPT is seen as a precondition for joint action planning with others and for solving social

[2] The ability of PT is inseparable from the so-called Theory of Mind (ToM, cf. Premack and Woodruff, 1978; Baron-Cohen et al., 1985). Both terms are sometimes used interchangeably (e.g., Birch et al., 2017), other researchers use a more complex ToM to emphasize the observer's insight that persons being observed may be in an individual state that differ from those of others. An observer, so to speak, can develop different ideas about what might be going on in the other person and weighs these ideas against each other in order to understand and to response accordingly (e.g., Harwood and Farrar, 2006). This comprehension of ToM, the possible weighing of different mental states, contributes only little to the present research question and is therefore neglected here.

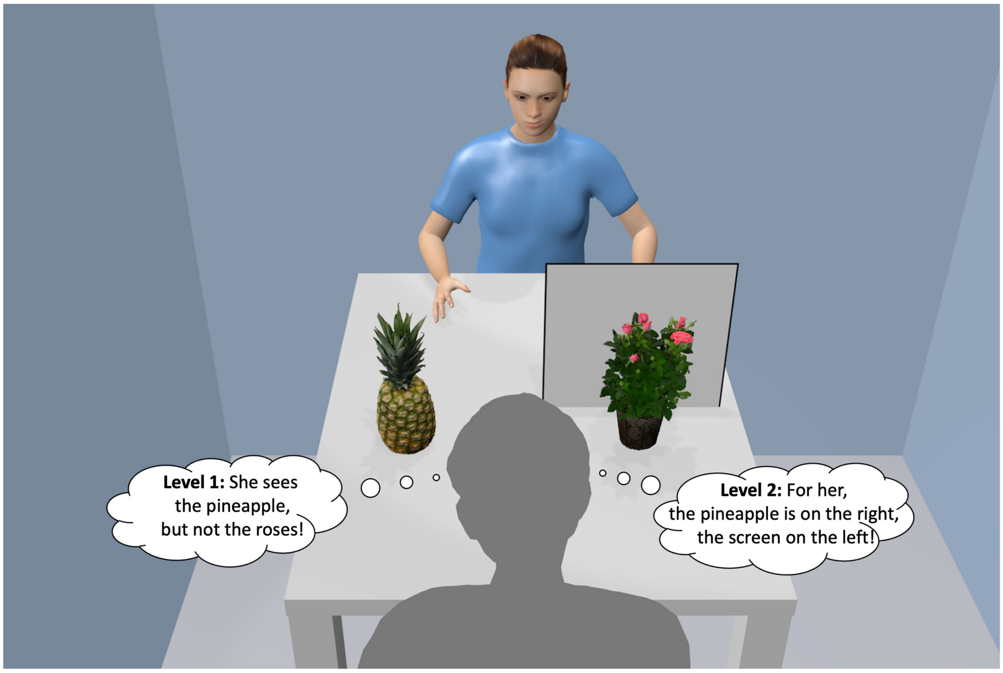

Level 1: She sees the pineapple, but not the roses!
Level 2: For her, the pineapple is on the right, the screen on the left!

FIGURE 1 | Level 1 and 2 visual-spatial perspective taking (VSPT) with regard to Flavell et al. (1981). ["Pineapple" (https://skfb.ly/6TQSO) and "Rose in a pot" (https://skfb.ly/6SDLR) by the sidekick are licensed under Creative Commons Attribution (http://creativecommons.org/licenses/by/4.0/)].

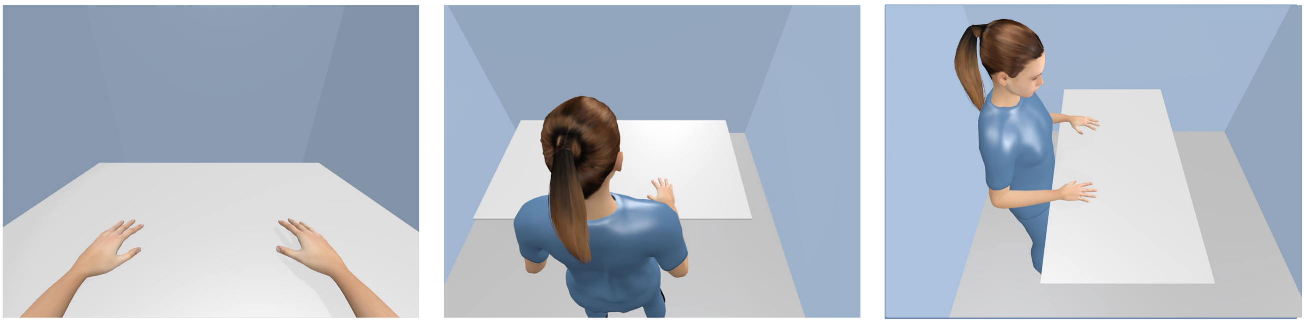


FIGURE 2 | The first person's visual perspective (**left panel**), the third person's visual perspective (here slightly lateral from above, **middle panel**), and the rotated visual perspective (here 90° clockwise rotated from the user's view**, right panel**).

tasks from the other's point of view (e.g., Freundlieb et al., 2017; Müsseler et al., 2019). Before getting into further details of level 2 VSPT, let is look at the different perspectives available for a user when dealing with an avatar in a virtual environment.

The First and Third Person Visual Perspective

The first person perspective is the view through the avatar's eyes (**Figure 2** left panel). The user sees the avatar's arms and hands as possible effectors and can sometimes look down to the avatar's legs (Pan and Steed, 2019), but the face, head, and back remain hidden (unless a mirror is in the virtual environment). Typical video games being played in the first person perspective are so-called first person shooters, such as Half-Life and the Call of Duty series. This perspective is often perceived as being close to reality, especially when the avatar's hands are the acting effectors in that virtual environment.

In a recent study, Arend and Müsseler (2021) showed that the presentation of avatar hands in the first person perspective facilitated responding to affording objects compared to a condition in which no hands were presented. This effect may be related to the finding outside of virtual environments that visual-spatial attention is preferentially directed to objects close to our real hands (near-hand effect, cf. Reed et al., 2006; Colman et al., 2017; Agauas et al., 2020). If a user has successfully

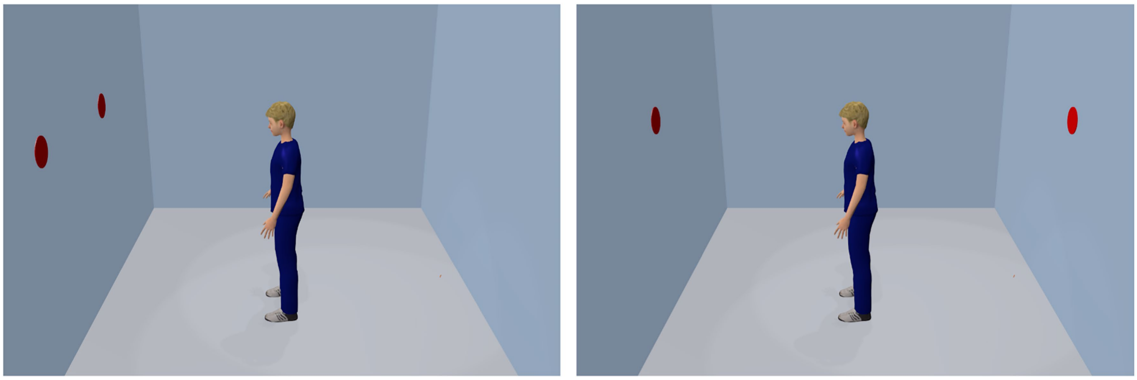


FIGURE 3 | The dot-perspective task of Samson et al. (2010). Participants responded to the number of dots on the display. Reaction times are typically facilitated when the participant sees the same number of dots as the avatar (**left panel**), compared to when they see a different number (**right panel**).

taken the avatar's perspective and sees the avatar's hands as their own hands, such effects should also be observable for the virtual hands, and this seems to be the case.

In the third person perspective, the user has the avatar's body in view, while the viewing direction is roughly maintained. So, the avatar is shown from behind, above, and/or slightly lateral (**Figure 2** middle panel).[3] Typical video games being played in the third person's perspective are Fortnite and the Witcher series.

Gorisse et al. (2017) carried out a study to compare the first with third person perspective. Their participants handled an avatar from either perspective in an immersive virtual environment. They found that the first person perspective enabled more accurate actions, while the third person perspective provides better spatial awareness (cf. the concept of self-location, Kilteni et al., 2012). Questionnaire data indicated the first person perspective as helpful to induce perceived ownership and to precise self-location. Kondo et al. (2018) also showed that the first person perspective was sufficient to induce perceived body ownership and that this impression was just as intense as the third person perspective with a whole-body avatar.

The Rotated Visual Perspective

The rotated visual perspective is a special type of the third person perspective, in which a person observes another individual viewing a scene from a completely different angle (**Figure 2** right panel). This situation characterizes primarily social encounters between humans, but it is also found in some video games with avatars (e.g., Grand Theft Auto 2 and games using isometric graphics or fixed camera positions).

Most of the research on VSPT has been conducted using this perspective, often with unanimated static avatars. An example is depicted in **Figure 3**, the so-called dot-perspective task introduced by Samson et al. (2010). The participant's task was to respond to the number of dots on a display. Reaction times were found to be facilitated when the participant sees the same number of dots as the avatar (left panel), compared to when they see a different number (right panel). This finding was interpreted as evidence for spontaneous perspective taking and is probably related to the tendency of humans to align their direction of gaze with one another (Driver et al., 1999; Frischen et al., 2007; Kunde et al., 2011).

A problem for the present research question is that the dot-perspective task and its findings may account for perspective taking (including that of an avatar), but less likely for avatar-self merging. This is because this task is purely receptive in nature and does not require acting from an avatar's perspective. We therefore favored the subsequent approach.

The Rotated Visual Perspective and User's Response Tendencies

The starting point for the following series of experiments was twofold (cf. Müsseler et al., 2019; Böffel and Müsseler, 2019b): First, a rotated visual perspective has the consequence that the spatial relations in a scene are different from the avatar's point of view and from the user's point of view. Second, cognitive psychology has shown that humans do possess predetermined response tendencies toward objects in space that sometimes facilitate one response and impede the other. The response tendencies of interest here are summarized under the label of spatial stimulus-response compatibility (for an overview see, e.g., Proctor and Vu, 2006). A typical finding in compatibility experiments is, for example, that a left (right) stimulus is responded faster and less error-prone with a compatible left (right) response than with an incompatible right (left) response.

In the present context, our aim was to confront participants with a situation that contained conflicting response tendencies from their own and their avatars' points of view and to observe which of the response tendencies dominated. If a user can become one with an avatar and act as if they are the avatar,

[3] Originally, labeling as first, second, and third person perspective comes from linguistic. First person is the I/we perspective, second person is the you perspective, and third person is the he/she/it/they perspective. However, the distinction of the second and third person perspectives does not make sense when considering the spatial relationships.

the response tendency from the avatar's point of view should prevail and override the one from the user's point of view.

The Avatar-Compatibility Task

Consider the following situation: A user controls the left and right hand of an avatar with left and right keystrokes. If the avatar is to grasp the handle of a pan lifter as shown in **Figure 4**, this suggests a right response from the avatar's point of view. However, the handle is oriented to the left from user's point of view, which should facilitate a left response. Thus, user's and avatar's perspective suggest different response tendencies and only if the user takes the perspective of the avatar, the right response should have an advantage. Or in other words, we hypothesized that users should neglect their own perspective when they become one with the avatar.

This was what we found in several studies and we refer to this compatibility effect from the avatar's point of view as the avatar-compatibility effect. In the experiments of Müsseler et al. (2019; see also Böffel and Müsseler, 2020a), participants should take the perspective of a rotated avatar and pressed ipsilateral or contralateral left-right keys in response to lateralized colored disks. We found consistently that compatibility effects were tied to the avatar's view but not to the participant's view. In other words, participants were able to perform compatible ipsilateral responses from the avatar's point of view faster and less error-prone than incompatible contralateral responses, even though from the participant's point of view the compatibility relationships were reversed. We interpret this finding as evidence that participants are able to implement their behavioral tendencies into the avatar, thereby neglecting their own perspective. Further note that compatibility findings (i.e., without an avatar) are usually very robust and can hardly be eliminated even by practice. It is therefore astonishing that the mere instruction to take the perspective of the avatar was able to turn the results into the opposite.

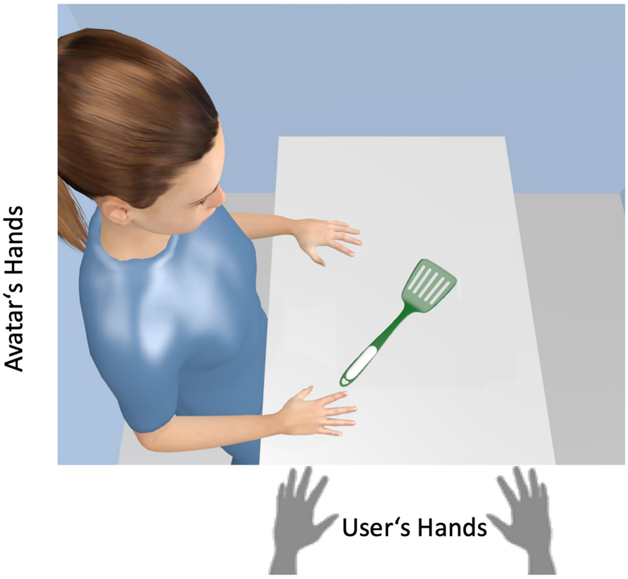


FIGURE 4 | The principle of the avatar-compatibility task. A user controls the left and right hand of an avatar with left and right keystrokes. If the avatar is to grasp the handle of the pan lifter, a right response from the avatar's point of view should be preferred (which required a right response of the user). However, the handle is oriented to the left from the user's point of view, which should facilitate a left response. Only if the user takes the perspective of the avatar, the right response should have the advantage. Our findings support consistently this assumption. ["Spatula" (https://skfb.ly/6QWQs) by Matthew is licensed under Creative Commons Attribution (http://creativecommons.org/licenses/by/4.0/). The color of the pan lifter was adjusted].

Böffel and Müsseler (2018) extended the finding by varying the degree of induced body ownership of the avatar *via* instruction. Half of the participants were informed to have complete control over an avatar (high-ownership condition), while the other half of the participants were informed that the avatar has its own will (low-ownership condition). Although the events on the screen were exactly the same in both conditions (for details of the experimental procedure, see Böffel and Müsseler, 2018), the results showed that the avatar-compatibility effect was more pronounced in the high-ownership condition than in the low-ownership condition. We attributed this to an increased avatar-self merging in the high-ownership condition compared with the low-ownership condition. This conclusion was supported by questionnaire data showing an increased body-ownership score in the high-ownership condition than in the low-ownership condition. The study demonstrated that body ownership and avatar-self merging rely on a person's interpretation of a situation that can be induced by the instruction.

The Avatar-Simon Task

While in the two previously mentioned studies the avatar could not be ignored to solve the task successfully, there is also evidence that the avatar's point of view is even adopted when it is in principle irrelevant for the task. A compatibility effect without an avatar, but task-irrelevant spatial positions is observed in the so-called Simon task (for an overview, see Hommel, 2011). Here, participants respond with the left-hand key to one color, for example, and with the right-hand key to another color that is presented on the left or right side of a display. Although stimulus position is task-irrelevant, spatially compatible conditions (e.g., left stimulus and left response) produce faster responses and fewer errors than spatially incompatible conditions (e.g., left stimulus and right response). Recent studies in our lab demonstrated that the Simon effect can also be observed when an avatar is added to the scene (**Figure 5**; Böffel and Müsseler, 2019b; von Salm-Hoogstraeten et al., 2020). By rotating the stimulus positions and the avatar by ±90° from the user's point of view, the stimulus does not contain spatial information on the left-right dimension from the user perspective, but only from the avatar perspective.

The results of the experiments indicated that actors take the avatar's perspective since they reacted in accordance with the Simon effect from the avatar's perspective (avatar-Simon effect; Böffel and Müsseler, 2019a,b, 2020b; von Salm-Hoogstraeten et al., 2020; von Salm-Hoogstraeten and Müsseler, 2021b). This finding also occurs spontaneously, that is, it is observed even when the participant is not instructed to take the avatar's perspective. However, when the avatar was replaced

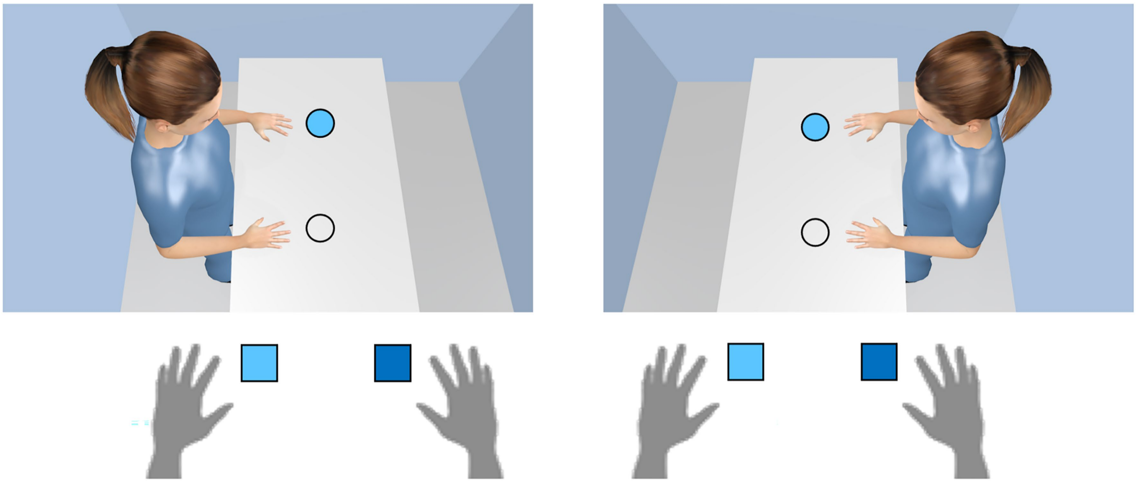


FIGURE 5 | The principle of the avatar-Simon task. Participant's task is to press on a light (dark) blue disk a left (right) key (here with light blue disk only). Disk positions are randomly assigned to the upper and lower position (here the upper position only). In the left panel, a left response is required, which corresponds to the avatar's left hand. In the right panel, a left response is also required, but it does not correspondent with the avatar's left hand, but its right hand. As a result, reaction times and fewer errors are observed with the avatar on the left side than with the avatar on the right side.

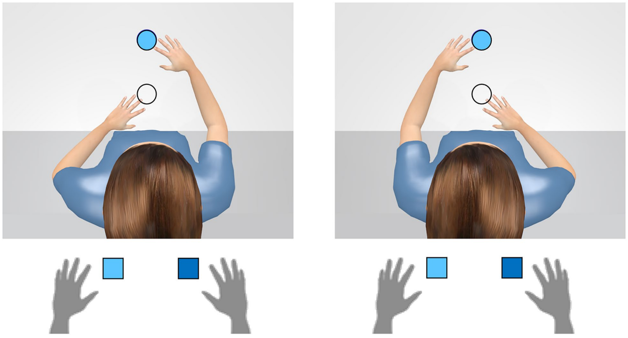


FIGURE 6 | The second scenario in the study of von Salm-Hoogstraeten et al. (2020). Participant took the first person perspective of the avatar and the avatar's right and left hand were now at the upper and lower stimulus position. The results showed pronounced avatar-Simon effects depending on the hand position of the avatar.

by a disk or an arc, the avatar-Simon effect disappeared (Böffel and Müsseler, 2019b). It is therefore obvious that not any simple object can trigger the effect and that a human-like character is beneficial. We will come back to this point below.

While the standard Simon effect (i.e., without an avatar) demonstrates that participants cannot ignore the position of a stimulus, the avatar-Simon effect shows additionally that they apparently cannot ignore also a (virtual) reference person either (for compatibility studies in social situations with human reference persons, see also Freundlieb et al., 2016, 2017).

"Seeing" the Avatar's Perspective vs. Referential Coding

Visual-spatial perspective taking is often understood as a process based on a visual–spatial representation created from another person's point of view. If the participants take the view of the avatar, they literally "see" the objects on the left or right side (e.g., Flavell, 1977; Costantini et al., 2011; Ward et al., 2019; for a critique of this view see Cole and Millett, 2019). Recent studies from our lab cast doubt on this simplification of the perspective-taking mechanism. von Salm-Hoogstraeten et al. (2020) compared two avatar scenarios: The first scenario was similar to the one illustrated in **Figure 5**. An avatar sat either to the left or to the right of a table and participants performed a Simon color-classification task to left-right stimuli from the viewpoint of the avatar. Note, that from the participants' point of view, the stimuli were arranged one above the other (i.e., with no spatial information on the horizontal dimension). The second scenario is illustrated in **Figure 6**. The participant took the first person perspective of the avatar and the avatar's right and left hand were now at the upper and lower stimulus position. In this scenario, only the avatar's hands formed the left and right relation to the stimulus positions. A perspective-created visual representation could only account for effects in the first scenario while the avatars' hands could produce a left-right frame of reference in both scenarios. The results showed pronounced avatar-Simon effects in both scenarios.

We interpreted this finding as evidence for the view that the avatar's position, and also the spatial positions of any other object in the scene, could be selected as a new spatial reference point from which the spatial relationships of the objects to each other could be redefined. That spatial coding of objects could arise in reference to other objects is an idea postulated by the referential coding account that was originally proposed to explain spatial compatibility effects in the standard Simon task (Hommel, 1993), and then was applied to the orthogonal compatibility task (Lippa, 1996; Cho and Proctor, 2005) and the object-based Simon task (Cho and Proctor, 2010; Arend and Müsseler, 2021). Recently, the referential coding account was also extended with regard to the joint Simon task (e.g., Dolk et al., 2013).

According to the referential coding account of perspective taking, the basic spatial map develop from the user's perspective, which, however, already contains all spatial relationships between objects in the visual space (cf. the visual sensory map of van der Heijden et al., 1999). Consequently, the user does not need to create a new visual-spatial map from the avatar's perspective but rather recodes the existing coordinates with regard to the new reference point. Thus, there may be little visual in visual perspective taking.

Generally, the recoding of objects within a new spatial reference frame is mostly investigated in terms of stimulus-coding, the mental representation of the objects and their positions. In a recent study (Böffel et al., 2020), we modified the avatar-Simon task by using centrally presented numbers as targets in order to remove the spatial variation of the stimuli. In these experiments, recoding the stimulus position could not be responsible for compatibility since the stimulus did not change its position. However, the avatar's movements could be recoded within the spatial reference frame and we still observed a compatibility effect, demonstrating that not only stimuli but also action effects are recoded from the avatar's point of view (Böffel et al., 2020). Therefore, the role of action effects and their spatial coding and interpretation seems to be crucial for avatar-based compatibility and was the topic of a series of further experiments.

VISUAL PERSPECTIVE TAKING WHEN CONTROLLING AVATAR'S MOVEMENTS

While the studies in the prior section used an avatar from whose perspective the user was supposed to act, the avatar itself did not perform the corresponding actions in all studies (e.g., not in Müsseler et al., 2019 and von Salm-Hoogstraeten et al., 2020).[4] It seems to be enough to imagine these movements (as in tool use, cf. Müsseler et al., 2014). However, it is indisputable that user movements that are synchronously and consistently mirrored in corresponding avatar movements increase perceived ownership and agency (e.g., Sanchez-Vives et al., 2010; Kilteni et al., 2012; Fox et al., 2015; Pfister et al., 2017; Kondo et al., 2018). The reason for this has already been noted in the Introduction: The ideomotor principle, as an integral part of the event-coding approach, allows to transform anticipated actions into executed actions (cf. James, 1890; Hommel et al., 2001; Kunde et al., 2004; Shin et al., 2010; Pfister, 2019). Furthermore, the comparison between anticipated and experienced outcomes contributes to who feels ownership of an action. Note that realizing these relationships is not a given from birth but is acquired in a developmental process in early childhood (e.g., Elsner and Adam, 2021). It also does not matter much where the action effects occur. In other words, whether action effects are anticipated in the proximal action space of the user (e.g., as tactile sensations at their hand) or in the distal space when a lamp is switched on or in the distal virtual space of the avatar depends alone on the user's intentions (cf. the findings with regard to tool use, e.g., Sutter et al., 2013).

Böffel and Müsseler (2019a) varied the participants' control over their avatar using the avatar-Simon task. In a full-control condition, the avatar consistently moved the left-right hand with the corresponding left-right keypress of the participant. In a less-control condition, the avatar moved a random hand instead, making the distal hand movements impossible to predict and effectively useless for action planning. The results confirmed our hypothesis that high control resulted in higher perceived body ownership and an increased avatar-Simon effect, providing evidence of increased avatar-self merging in both self-report and behavioral data (see also Ma and Hommel, 2015).

Consistent action effects at the avatar also allow the user to differentiate their avatar from other characters (which are controlled by another user or by the computer program). Self-other distinction is an important requirement for successful interactions in real and virtual environments (e.g., Mattan et al., 2016). Only the identification of one's own avatar and the differentiation from others enables successful action. This can be achieved by consistent feedback of the anticipated action effects at the own avatar. von Salm-Hoogstraeten and Müsseler (2021b) showed that users preferred to take the perspective of the avatar that consistently mirrored their actions, even though another virtual character took a similar perspective. The study also showed that perspective taking is not that spontaneous, as sometimes assumed (cf. Samson et al., 2010; Freundlieb et al., 2016, 2017). Instead, perspective taking is likely to benefit from action-based and thereby top-down controlled processes.

Besides the consistency of action effects, the synchronicity and movement correspondence of action effects of the avatar is likely to be conducive to avatar-self merging. Although not examined in a study with an avatar, it is likely that the actor no longer experiences themselves as the originator of an action, when the action effect is presented too early (e.g., before the user's action) or too late (cf. Haering and Kiesel, 2015; Dignath and Janczyk, 2017). Similarly, performance decreases if action effects are durationally or spatially not in correspondence with the participant's movements, e.g., when a short keystroke is transferred into a long keystroke or a right movement into a left movement (or vice versa; Pfister et al., 2017; Liesner et al., 2020).

As with the rubber-hand illusion, attention should also be paid to corresponding hand-hand postures (cf. Costantini and Haggard, 2007). In yet unpublished experiments in our lab, we were able to show that both the avatar-compatibility effect and the avatar-Simon effect disappeared when either the avatar or the user crossed their hands. This was despite the fact that hand-hand correspondence still applied, that is, a left (right) button press resulted in a left (right) action effect at the corresponding hand of the avatar. Only when both pairs of hands, the user's and the avatar's, were crossed, the effects re-appeared in both objective and subjective measures (Müsseler, 2019). In summary,

[4] In these studies, a static unanimated avatar was used to clearly attribute the findings to perspective taking and not to the appearance of anticipated action effects at the avatar (see below and Kunde, 2001; Müsseler and Skottke, 2011).

appropriate action effects at the avatar (with regard to consistency, synchronicity, correspondence, and posture) not only facilitate self-merging with the avatar, they also contribute essentially to self-other distinction within the virtual environment.

VISUAL-SPATIAL PERSPECTIVE TAKING AS A SOCIAL ABILITY

There is an ongoing debate about whether the ability of VSPT emerges exclusively in social interpersonal contexts (referring to the more cognitively demanding level 2 VSPT; Flavell et al., 1981). Can one also take the perspective of a (humanoid) character or even an object? Since the seminal paper of Shepard and Metzler (1971), the ability to mentally rotate an object is undisputed. However, note that in VSPT, humans perform a mental self-rotation in order to take the perspective of others. This makes perspective taking with (humanoid) characters and mental rotation with objects dissociable (e.g., Zacks and Michelon, 2005; Kessler and Thomson, 2010). Still, Hegarty and Waller (2004) reported that both abilities are highly correlated, which could indicate that perspective taking is not tied to human or humanoid characters. Accordingly, we observed the avatar-Simon effect also with a headless robot that could hardly be described as humanoid (von Salm-Hoogstraeten and Müsseler, 2021a). However, the robot had two arms and perhaps that was enough to yield a humanoid appearance. At least the two arms could have specified the direction of perspective taking, which is normally determined by the gaze direction or head orientation of the observed character. This in turn strengthens the social view of perspective taking, because objects usually do not have this orientation.

Evidence emphasizing the social aspect of VSPT has been recently reported in a study by Ward et al. (2019). Their participants judged normal or mirrored letters (e.g., an R or an Я) shown with various rotation angles on a flat table. Either only the table was presented or an avatar sat to its left or right or a lamp directed toward the letters was placed at the same position as the avatar. The authors observed lower response times with low rotation angles of the participants to the letters compared to larger angles. However, lower response times were also found when the rotation angles were low with regard to the avatar, although, then, the angle with regard to the participants was high. Most importantly in the present context, no such effects were observed with the lamp presented instead of the avatar. This is in line with our observations that the avatar-Simon effect disappeared when a disk or an arc was presented instead of the avatar (Böffel and Müsseler, 2019b).

To a last example focusing on the social aspect in virtual environments: In the experiments of Bönsch et al. (2018, 2020), users controlled an avatar in space in the first person perspective, which was approached by either a happy-looking or angry-looking virtual character. Users preferred to be at a greater distance from or walk past the angry-looking character than the happy-looking character. These results show that the regularities that apply in human-human interaction are also adopted in virtual environments. Whether this can be interpreted beyond doubt as evidence for avatar-self merging is debatable, but at least maintaining these regularities in virtual environments should facilitate it.

CONCLUSION

In this paper, we started with the event-coding approach of the self (Hommel, 2018, 2021) and showed that self-avatar overlap is predestined to give rise to avatar-self merging, mainly due to the transfer of the user's motor activities into corresponding avatar activities. For successful avatar-self merging, it seems essential to us that the virtual environment opens up possible actions for the user to realize their intentions. Whether action control is achieved in a real environment or an artificial one is not decisive for the self.

In our experiments, users were confronted with situations that contained conflicting response tendencies from their own and their avatars' points of view. The results revealed that users often overrode their own response tendencies and acted as if they were the avatar. As a rule, this observation was accompanied by increased scores in perceived ownership and agency (Böffel and Müsseler, 2018, 2019a), suggesting avatar-self merging. The procedure of our experiments could be applied to a variety of other response tendencies that are known in cognitive psychology.

For example, so far, we have dealt almost exclusively with spatial stimulus-response compatibilities, that is, both stimuli and responses exhibited a critical spatial position (but see Böffel et al., 2020). However, there are also stimuli that trigger response tendencies regardless of their spatial position. For instance, the presentation of a baby photo usually produces an approach behavior, whereas the photo of a violent scene produces an avoidance behavior (e.g., gathered with a speeded joystick response, Eder et al., 2012). If an avatar is added to the scene, from whose point of view the photos are to be judged, the experimenter can again create a discrepancy from the user and avatar point of view and examine which response tendency dominates. Further, it would be intriguing to examine whether the user also adopts social attitudes of an avatar, which are associated with its ethnicity, its gender, or—more general—its group affiliation. Again, to clearly interpret the results, it would be important to ensure an experimental setup with a discrepancy between the user's attitudes and the avatar's affiliation.

Following the event-coding approach, avatar-self merging is not seen as an all-or-nothing process, but rather as a process to which different features may or may not contribute. As various studies have shown, the human information-processing system is flexible enough to adapt its behavior not only to various real-world environments but also to novel artificial virtual ones. As a prerequisite for avatar-self merging, we consider the user's ability to successfully take the perspective of an avatar in affective, cognitive, and visual–spatial terms. However, this is not to say that these factors are adopted in their entirety. This remains an empirical question.

In addition to the cognitive aspects, the extent of avatar-self merging is of course also determined by the technical implementations of the virtual environment. The more immersive a virtual environment is, the more likely our senses are to experience an environment as "real," and the more pronounced avatar-self merging is likely to be. However, immersion also means that the senses important for action planning and action execution are implemented, that is, the efferent mechanisms triggering an action and the afferent mechanisms controlling them. In this context, it should also be pointed out that most (action) events in our natural environment can be experienced in a multisensorial manner (i.e., visual, auditory, tactile, and/or proprioceptive). This is often missing in the virtual applications.

Even if we succeeded in realizing all these components in an immersive environment, the problem of sensorimotor transformation would remain. It consists in transforming a proximal movement (e.g., a user's keypress) into a non-corresponding distal movement (e.g., a movement of the entire hand including the arm of an avatar; cf. this problem in tool use, Sutter et al., 2013). Thus, this transformation rarely follows a 1:1 rule but is, for example, longer or shorter, amplified, or reduced in force, and this not necessarily in a linear manner. Acquisition and execution of distal movements in the presence of sensorimotor transformations are challenging for any user. That is the bad news. The good news is that the human users have the ability to acquire these transformations (although sometimes with a lot of practice) and then can act accordingly. As a consequence, avatar-self merging needs time and occurs only when the users have sufficiently internalized the transformation rule between proximal and distal action effects.

AUTHOR CONTRIBUTIONS

JM: conceptualization, writing—original draft, funding acquisition, project administration, and supervision. SS-H: conceptualization, visualization, and writing—review and editing. CB: conceptualization and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; project number MU 1298/11-1) and was associated with the DFG Priority Program "The Active Self" (DFG SPP 2134).

REFERENCES

Agauas, S. J., Jacoby, M., and Thomas, L. E. (2020). Near-hand effects are robust: three OSF pre-registered replications of visual biases in perihand space. *Vis. Cogn.* 28, 192–204. doi: 10.1080/13506285.2020.1751763

Arend, M., and Müsseler, J. (2021). Object affordances from the perspective of an avatar. *Conscious. Cogn.* 92:103133. doi: 10.1016/j.concog.2021.103133

Ash, E. (2016). Priming or proteus effect? Examining the effects of avatar race on in-game behavior and post-play aggressive cognition and affect in video games. *Games Cult.* 11, 422–440. doi: 10.1177/1555412014568870

Banakou, D., Groten, R., and Slater, M. (2013). Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proc. Natl. Acad. Sci. U. S. A.* 110, 12846–12851. doi: 10.1073/pnas.1306779110

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Beaudoin, M., Barra, J., Dupraz, L., Mollier-Sabet, P., and Guerraz, M. (2020). The impact of embodying an "elderly" body avatar on motor imagery. *Exp. Brain Res.* 238, 1467–1478. doi: 10.1007/s00221-020-05828-5

Bekrater-Bodmann, R. (2020). Perceptual correlates of successful body-prosthesis interaction in lower limb amputees: psychometric characterisation and development of the prosthesis embodiment scale. *Sci. Rep.* 10:14204. doi: 10.1038/s41598-020-70828-y

Birch, S. A. J., Li, V., Haddock, T., Ghrear, S. E., Brosseau-Liard, P., Baimel, A., et al. (2017). Perspectives on perspective taking: how children think about the minds of others. *Adv. Child Dev. Behav.* 52, 185–226. doi: 10.1016/bs.acdb.2016.10.005

Böffel, C. (2021). Avatar-induced compatibilities and the concept of avatar-self merging. doctoral dissertation. RWTH Aachen University.

Böffel, C., Herbst, C., Lindemann, O., and Müsseler, J. (2020). Spatial–numerical associations in the presence of an avatar. *Psychol. Res.* 85, 2588–2598. doi: 10.1007/s00426-020-01424-y

Böffel, C., and Müsseler, J. (2018). Perceived ownership of avatars influences visual perspective taking. *Front. Psychol.* 9:743. doi: 10.3389/fpsyg.2018.00743

Böffel, C., and Müsseler, J. (2019a). Action effect consistency and body ownership in the avatar-Simon task. *PLoS One* 14:e0220817. doi: 10.1371/journal.pone.0220817

Böffel, C., and Müsseler, J. (2019b). Visual perspective taking for avatars in a Simon task. *Atten. Percept. Psychophysiol.* 81, 158–172. doi: 10.3758/s13414-018-1573-0

Böffel, C., and Müsseler, J. (2020a). No evidence for automatic response activation with target onset in the avatar-compatibility task. *Mem. Cogn.* 48, 1249–1262. doi: 10.3758/s13421-020-01052-2

Böffel, C., and Müsseler, J. (2020b). Taking time to take perspective? Rapidly changing reference frames in the avatar-Simon task. *Acta Psychol.* 204:103005. doi: 10.1016/j.actpsy.2020.103005

Bönsch, A., Radke, S., Ehret, J., Habel, U., and Kuhlen, T. W. (2020). "The impact of a virtual agent's non-verbal emotional expression on a user's personal space preferences." in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*, October 20-22, 2020.

Bönsch, A., Radke, S., Overath, H., Asche, L. M., Wendt, J., Vierjahn, T., et al. (2018). "Social VR: how personal space is affected by virtual agents' emotions." in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, March 22-26, 2020.

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature* 391:756. doi: 10.1038/35784

Cameron, J. (2009). *Avatar*. Los Angeles, CA: 20th Century Fox.

Cho, Y. S., and Proctor, R. W. (2005). Representing response position relative to display location: influence on orthogonal stimulus-response compatibility. *Q. J. Exp. Psychol. A* 58, 839–864. doi: 10.1080/02724980443000359

Cho, D. T., and Proctor, R. W. (2010). The object-based Simon effect: grasping affordance or relative location of the graspable part? *J. Exp. Psychol. Hum. Percept. Perform.* 36, 853–861. doi: 10.1037/a0019328

Cole, G. G., and Millett, A. C. (2019). The closing of the theory of mind: a critique of perspective-taking. *Psychon. Bull. Rev.* 26, 1787–1802. doi: 10.3758/s13423-019-01657-y

Colman, H. A., Remington, R. W., and Kritikos, A. (2017). Handedness and graspability modify shifts of visuospatial attention to near-hand objects. *PLoS One* 12:e0170542. doi: 10.1371/journal.pone.0170542

Costantini, M., Committieri, G., and Sinigaglia, C. (2011). Ready both to your and to my hands: mapping the action space of others. *PLoS One* 6:e17923. doi: 10.1371/journal.pone.0017923

Costantini, M., and Haggard, P. (2007). The rubber hand illusion: sensitivity and reference frame for body ownership. *Conscious. Cogn.* 16, 229–240. doi: 10.1016/j.concog.2007.01.001

de Haan, S., and de Bruin, L. (2010). Reconstructing the minimal self, or how to make sense of agency and ownership. *Phenomenol. Cogn. Sci.* 9, 373–396. doi: 10.1007/s11097-009-9148-0

Dignath, D., and Janczyk, M. (2017). Anticipation of delayed action-effects: learning when an effect occurs, without knowing what this effect will be. *Psychol. Res.* 81, 1072–1083. doi: 10.1007/s00426-016-0797-7

Dolk, T., Hommel, B., Prinz, W., and Liepelt, R. (2013). The (not so) social Simon effect: a referential coding account. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 1248–1260. doi: 10.1037/a0031031

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., and Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Vis. Cogn.* 6, 509–540. doi: 10.1080/135062899394920

Dunn, R. A., and Guadagno, R. E. (2012). My avatar and me – gender and personality predictors of avatar-self discrepancy. *Comput. Hum. Behav.* 28, 97–106. doi: 10.1016/j.chb.2011.08.015

Dunn, R. A., and Guadagno, R. (2019). Who are you online? A study of gender, race, and gaming experience and context on avatar self-representation. *Int. J. Cyber Behav. Psychol. Learn.* 9, 15–31. doi: 10.4018/IJCBPL.2019070102

Eder, A. B., Müsseler, J., and Hommel, B. (2012). The structure of affective action representations: temporal binding of affective response codes. *Psychol. Res.* 76, 111–118. doi: 10.1007/s00426-011-0327-6

Elsner, B., and Adam, M. (2021). Infants' goal prediction for simple action events: The role of experience and agency cues. *Top. Cogn. Sci.* 13, 45–62. doi: 10.1111/tops.12494

Flavell, J. H. (1977). The development of knowledge about visual perception. *Neb. Symp. Motiv.* 25, 43–76

Flavell, J. H., Everett, B. A., Croft, K., and Flavell, E. R. (1981). Young children's knowledge about visual-perception: further evidence for the Level-1–Level-2 distinction. *Dev. Psychol.* 17, 99–103. doi: 10.1037/0012-1649.17.1.99

Fox, J., Ahn, S. J. G., Janssen, J. H., Yeykelis, L., Segovia, K. Y., and Bailenson, J. N. (2015). Avatars versus agents: a meta-analysis quantifying the effect of agency on social influence. *Hum. Comput. Interact.* 30, 401–432. doi: 10.1080/07370024.2014.921494

Fox, J., and Bailenson, J. N. (2009). Virtual self-modeling: the effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychol.* 12, 1–25. doi: 10.1080/15213260802669474

Freundlieb, M., Kovács, Á. M., and Sebanz, N. (2016). When do humans spontaneously adopt another's visuospatial perspective? *J. Exp. Psychol. Hum. Percept. Perform.* 42, 401–412. doi: 10.1037/xhp0000153

Freundlieb, M., Sebanz, N., and Kovács, Á. M. (2017). Out of your sight, out of my mind: knowledge about another person's visual access modulates spontaneous visuospatial perspective-taking. *J. Exp. Psychol. Hum. Percept. Perform.* 43, 1065–1072. doi: 10.1037/xhp0000379

Fribourg, R., Argelaguet, F., Lécuyer, A., and Hoyet, L. (2020). Avatar and sense of embodiment: studying the relative preference between appearance, control and point of view. *IEEE Trans. Vis. Comput. Graph.* 26, 2062–2072. doi: 10.1109/TVCG.2020.2973077

Frischen, A., Bayliss, A., and Tipper, S. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull.* 133, 694–724. doi: 10.1037/0033-2909.133.4.694

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/s1364-6613(99)01417-5

Gallese, V., and Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia* 48, 746–755. doi: 10.1016/j.neuropsychologia.2009.09.038

Gorisse, G., Christmann, O., Amato, E. A., and Richir, S. (2017). First- and third-person perspectives in immersive virtual environments: presence and performance analysis of embodied users. *Front. Robot. AI* 4:33. doi: 10.3389/frobt.2017.00033

Haering, C., and Kiesel, A. (2015). Was it me when it happened too early? Experience of delayed effects shapes sense of agency. *Cognition* 136, 38–42. doi: 10.1016/j.cognition.2014.11.012

Harwood, M. D., and Farrar, M. J. (2006). Conflicting emotions: the connection between affective perspective taking and theory of mind. *Br. J. Dev. Psychol.* 24, 401–418. doi: 10.1348/026151005X50302

Hefner, D., Klimmt, C., and Vorderer, P. (2007). "Identification with the player character as determinant of video game enjoyment," in *Entertainment Computing – ICEC 2007. Lecture Notes in Computer Science, Vol. 4740*. eds. L. Ma, M. Rauterberg and R. Nakatsu (Berlin, Heidelberg: Springer)

Hegarty, M., and Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence* 32, 175–191. doi: 10.1016/j.intell.2003.12.001

Hershfield, H. E., Goldstein, D. G., Sharpe, W. F., Fox, J., Yeykelis, L., Carstensen, L. L., et al. (2011). Increasing saving behavior through age-progressed renderings of the future self. *J. Mark. Res.* 48, S23–S37. doi: 10.1509/jmkr.48.SPL.S23

Hommel, B. (1993). Inverting the Simon effect by intention: determinants of direction and extent of effects of irrelevant spatial information. *Psychol. Res.* 55, 270–279. doi: 10.1007/BF00419687

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychol.* 136, 189–202. doi: 10.1016/j.actpsy.2010.04.011

Hommel, B. (2018). Representing oneself and others: an event-coding approach. *Exp. Psychol.* 65, 323–331. doi: 10.1027/1618-3169/a000433

Hommel, B. (2021). The me-file: an event-coding approach to self-representation. *Front. Psychol.* 12:698778. doi: 10.3389/fpsyg.2021.698778

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding: a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–878. doi: 10.1017/S0140525X01000103

James, W. (1890). *The Principles of Psychology. Vol. 2*. Cambridge, MA: Harvard University Press.

Kessler, K., and Thomson, L. A. (2010). The embodied nature of spatial perspective taking: embodied transformation versus sensorimotor interference. *Cognition* 114, 72–88. doi: 10.1016/j.cognition.2009.08.015

Kilteni, K., Groten, R., and Slater, M. (2012). The sense of embodiment in virtual reality. *Presence Teleop. Virt.* 21, 373–387. doi: 10.1162/PRES_a_00124

Kilteni, K., Maselli, A., Kording, K. P., and Slater, M. (2015). Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Front. Hum. Neurosci.* 9:141. doi: 10.3389/fnhum.2015.00141

Kondo, R., Sugimoto, M., Minamizawa, K., Hoshi, T., Inami, M., and Kitazaki, M. (2018). Illusory body ownership of an invisible body interpolated between virtual hands and feet via visual-motor synchronicity. *Sci. Rep.* 8:7541. doi: 10.1038/s41598-018-25951-2

Kunde, W. (2001). Response-effect compatibility in manual choice reaction tasks. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 387–394. doi: 10.1037/0096-1523.27.2.387

Kunde, W., Koch, I., and Hoffmann, J. (2004). Anticipated action effects affect the selection, initiation, and execution of actions. *Q. J. Exp. Psychol. A* 57, 87–106. doi: 10.1080/02724980343000143

Kunde, W., Skirde, S., and Weigelt, M. (2011). Trust my face: cognitive factors of head fakes in sports. *J. Exp. Psychol. Appl.* 17, 110–127. doi: 10.1037/a0023756

Latoschik, M. E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., and Botsch, M. (2017). "The effect of avatar realism in immersive social virtual realities." in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, November 8-10, 2017; 1–10.

Liesner, M., Kirsch, W., and Kunde, W. (2020). The interplay of predictive and postdictive components of experienced selfhood. *Conscious. Cogn.* 77:102850. doi: 10.1016/j.concog.2019.102850

Lippa, Y. (1996). A referential coding explanation for compatibility effects of physically orthogonal stimulus and response dimensions. *Q. J. Exp. Psychol.* 49, 950–971. doi: 10.1080/713755676

Lugrin, J. L., Landeck, M., and Latoschik, M. E. (2015). Avatar embodiment realism and virtual fitness training. *IEEE Virtual Real.* 225–226. doi: 10.1109/VR.2015.7223377

Ma, K., and Hommel, B. (2015). The role of agency for perceived ownership in the virtual hand illusion. *Conscious. Cogn.* 36, 277–288. doi: 10.1016/j.concog.2015.07.008

Ma, K., Hommel, B., and Cheng, H. (2019). The roles of consistency and exclusivity in perceiving body ownership and agency. *Psychol. Res.* 83, 175–184. doi: 10.1007/s00426-018-0978-7

Ma, K., Qu, J., Yang, L., Zhao, W., and Hommel, B. (2021). Explicit and implicit measures of body ownership and agency: affected by the same manipulations and yet independent. *Exp. Brain Res.* 239, 2159–2170. doi: 10.1007/s00221-021-06125-5

Mattan, B. D., Rotshtein, P., and Quinn, K. A. (2016). Empathy and visual perspective-taking performance. *Cogn. Neurosci.* 7, 170–181. doi: 10.1080/17588928.2015.1085372

Müsseler, J. (2019). Testing the limits when taking avatar's perspective: deviating hand positions between actor and avatar. *Abstr. Psychon. Soc.* 24:267.

Müsseler, J., Ruhland, L., and Böffel, C. (2019). Reversed effect of spatial compatibility when taking avatar's perspective. *Q. J. Exp. Psychol.* 72, 1539–1549. doi: 10.1177/1747021818799240

Müsseler, J., and Skottke, E.-M. (2011). Compatibility relationships with simple lever tools. *Hum. Factors* 53, 383–390. doi: 10.1177/0018720811408599

Müsseler, J., Wühr, P., and Ziessler, M. (2014). Using tools with real and imagined tool movements. *Front. Psychol.* 5:515. doi: 10.3389/fpsyg.2014.00515

Pan, X., and Hamilton, A. F. C. (2018). Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape. *Br. J. Psychol.* 109, 395–417. doi: 10.1111/bjop.12290

Pan, Y., and Steed, A. (2019). How foot tracking matters: the impact of an animated self-avatar on interaction, embodiment and presence in shared virtual environments. *Front. Robot. AI* 6:104. doi: 10.3389/frobt.2019.00104

Peck, T. C., and Gonzalez-Franco, M. (2021). Avatar embodiment. Towards a standardized questionnaire. *Front. Virtual Real.* 1:575943. doi: 10.3389/frvir.2020.575943

Peña, J., Hancock, J. T., and Merola, N. A. (2009). The priming effects of avatars in virtual settings. *Commun. Res.* 36, 838–856. doi: 10.1177/0093650209346802

Pfister, R. (2019). Effect-based action control with body-related effects: implications for empirical approaches to ideomotor action control. *Psychol. Rev.* 126, 153–161. doi: 10.1037/rev0000140

Pfister, R., Weller, L., Dignath, D., and Kunde, W. (2017). What or when? The impact of anticipated social action effects is driven by action-effect compatibility, not delay. *Atten. Percept. Psychophysiol.* 79, 2132–2142. doi: 10.3758/s13414-017-1371-0

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/S0140525X00076512

Proctor, R. W., and Vu, K.-P. L. (2006). *Stimulus-Response Compatibility Principles: Data, Theory, and Application*. Boca Raton, FL: CRC Press.

Quintard, V., Jouffe, S., Hommel, B., and Bouquet, C. A. (2021). Embodied self-other overlap in romantic love: a review and integrative perspective. *Psychol. Res.* 85, 899–914. doi: 10.1007/s00426-020-01301-8

Ratan, R., Beyea, D., Li, B. J., and Graciano, L. (2020). Avatar characteristics induce users' behavioral conformity with small-to-medium effect sizes: a meta-analysis of the proteus effect. *Media Psychol.* 23, 651–675. doi: 10.1080/15213269.2019.1623698

Reed, C. L., Grubb, J. D., and Steele, C. (2006). Hands up: attentional prioritization of space near the hand. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 166–177. doi: 10.1037/0096-1523.32.1.166

Reinhard, R., Shah, K. G., Faust-Christmann, C. A., and Lachmann, T. (2020). Acting your avatar's age: effects of virtual reality avatar embodiment on real life walking speed. *Media Psychol.* 23, 293–315. doi: 10.1080/15213269.2019.1598435

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., and Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1255–1266. doi: 10.1037/a0018729

Sanchez-Vives, M. V., Spanlang, B., Frisoli, A., Bergamasco, M., and Slater, M. (2010). Virtual hand illusion induced by visuomotor correlations. *PLoS One* 5:e10381. doi: 10.1371/journal.pone.0010381

Shepard, R. N., and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science* 171, 701–703. doi: 10.1126/science.171.3972.701

Shin, Y. K., Proctor, R. W., and Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychol. Bull.* 136, 943–974. doi: 10.1037/a0020541

Steins, G., and Wicklund, R. A. (1993). Zum Konzept der Perspektivenübernahme: Ein kritischer Überblick [The concept of perspective-taking: A critical overview]. *Psychol. Rundsch.* 44, 226–239.

Sutter, C., Sülzenbrück, S., Rieger, M., and Müsseler, J. (2013). Limitations of distal effect anticipation when using tools. *New Ideas Psychol.* 31, 247–257. doi: 10.1016/j.newideapsych.2012.12.001

van der Heijden, A. H. C., Müsseler, J., and Bridgeman, B. (1999). On the perception of position. *Adv. Psychol.* 129, 19–37. doi: 10.1016/S0166-4115(99)80005-3

Verschoor, S. A., and Hommel, B. (2017). Self-by-doing: the role of action for self-acquisition. *Soc. Cogn.* 35, 127–145. doi: 10.1521/soco.2017.35.2.127

von Salm-Hoogstraeten, S., Bolzius, K., and Müsseler, J. (2020). Seeing the world through the eyes of an avatar? Comparing perspective taking and referential coding. *J. Exp. Psychol. Hum. Percept. Perform.* 46, 264–273. doi: 10.1037/xhp0000711

von Salm-Hoogstraeten, S., and Müsseler, J. (2021a). Human cognition in interaction with robots: taking the robot's perspective into account. *Hum. Factors* 63, 1396–1407. doi: 10.1177/0018720820933764

von Salm-Hoogstraeten, S., and Müsseler, J. (2021b). Perspective taking while interacting with a self-controlled or independently-acting avatar. *Comput. Hum. Behav.* 118:106698. doi: 10.1016/j.chb.2021.106698

Ward, E., Ganis, G., and Bach, P. (2019). Spontaneous vicarious perception of the content of another's visual perspective. *Curr. Biol.* 29, 874.e4–880.e4. doi: 10.1016/j.cub.2019.01.046

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

Wolf, D., and Henley, A. J. (eds.) (2017). "Use gravatar to display user's avatars with posts" in *Java EE Web Application Primer* (Berkeley, CA: Apress).

Yee, N., and Bailenson, J. (2007). The proteus effect: the effect of transformed self-representation on behavior. *Hum. Commun. Res.* 33, 271–290. doi: 10.1111/j.1468-2958.2007.00299.x

Yoon, G., and Vargas, P. T. (2014). Know thy avatar: the unintended effect of virtual-self representation on behavior. *Psychol. Sci.* 25, 1043–1045. doi: 10.1177/0956797613519271

Zacks, J. M., and Michelon, P. (2005). Transformations of visuospatial images. *Behav. Cogn. Neurosci. Rev.* 4, 96–118. doi: 10.1177/1534582305281085

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

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

ORIGINAL RESEARCH
published: 08 April 2022
doi: 10.3389/fpsyg.2022.717669

Habituation and Dishabituation in Motor Behavior: Experiment and Neural Dynamic Model

Sophie Aerdker[1]*, *Jing Feng*[2] *and Gregor Schöner*[1]

[1] *Institute for Neural Computation, Ruhr University Bochum, Bochum, Germany,* [2] *Motion Analysis Center, Shriners Hospitals for Children, Portland, OR, United States*

Does motor behavior early in development have the same signatures of habituation, dishabituation, and Spencer-Thompson dishabituation known from infant perception and cognition? And do these signatures explain the choice preferences in A not B motor decision tasks? We provide new empirical evidence that gives an affirmative answer to the first question together with a unified neural dynamic model that gives an affirmative answer to the second question.In the perceptual and cognitive domains, habituation is the weakening of an orientation response to a stimulus over perceptual experience. Switching to a novel stimulus leads to dishabituation, the re-establishment of the orientation response. In Spencer-Thompson dishabituation, the renewed orientation response transfers to the original (familiar) stimulus. The change in orientation responses over perceptual experience explains infants' behavior in preferential looking tasks: Familiarity preference (looking longer at familiar than at novel stimuli) early during exposure and novelty preference (looking longer at novel than at familiar stimuli) late during exposure. In the motor domain, perseveration in the A not B task could be interpreted as a form of familiarity preference. There are hints that this preference reverses after enough experience with the familiar movement. We provide a unified account for habituation and patterns of preferential selection in which neural dynamic fields generate perceptual or motor representations. The build-up of activation in excitatory fields leads to familiarity preference, the build-up of activation in inhibitory fields leads to novelty preference. We show that the model accounts for the new experimental evidence for motor habituation, but is also compatible with earlier accounts for perceptual habituation and motor perseveration. We discuss how excitatory and inhibitory memory traces may regulate exploration and exploitation for both orientation to objects and motor behaviors.

Keywords: habituation, perseveration, neural dynamic model, Dynamic Field Theory, exploration-exploitation

OPEN ACCESS

Edited by:
Dongheui Lee,
Technical University of Munich,
Germany

Reviewed by:
Martin V. Butz,
University of Tübingen, Germany
Efthimios M. C. Skoulakis,
Alexander Fleming Biomedical
Sciences Research Center, Greece

****Correspondence:***
Sophie Aerdker
sophie.aerdker@ini.rub.de

Specialty section:
This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

Received: *31 May 2021*
Accepted: *21 February 2022*
Published: *08 April 2022*

Citation:
Aerdker S, Feng J and Schöner G
(2022) Habituation and Dishabituation
in Motor Behavior: Experiment and
Neural Dynamic Model.
Front. Psychol. 13:717669.
doi: 10.3389/fpsyg.2022.717669

1. INTRODUCTION

Most behavior is directed at objects in the world that are perceived based on sensory information. Once a particular object has been selected as the target of an action, other objects may effectively become distractors. A selected action must be stabilized against competing actions directed at these other objects. In the development of object-directed action, perseverative reaching may be viewed as a signature of such stabilization (Smith et al., 1999; Thelen et al., 2001). In the classical A not B paradigm (Wellman et al., 1986), infants repeatedly reach for a toy that is hidden at one of two locations, typically two troughs cut out from a box and covered by lids. On each trial, the infant watches as the experimenter hides the toy at the A location and then, after a short delay, pushes the box into the infant's reaching space. After the infant reaches, and typically retrieves the toy, the experimenter gently wrings the toy out of the infant's hand, pulls the box back to its starting position and starts another trial. After six such "A trials," the toy is next hidden at the B location. Young infants (from around 7 to 10 months of age) then typically perseverate, reaching again for the A location rather than retrieving the toy at the B location. In a sense, they stabilize the reach to A, suppressing the distractor cue to B. Older infants do not make this perseverative error. They are able to follow the cue and switch to the B location.

Habituation is commonly observed in paradigms that probe infant perception and cognition (Colombo and Mitchell, 2009). In a typical visual habituation paradigm a salient visual stimulus is presented to an infant against a nondescript background. Infants' orientation response is measured through "looking time" (total duration of fixation on the stimulus) or by physiological measures such as increased heart rate or sucking frequency. Presentation is repeated, often in a manner that depends on the infants response. A trial starts once the infant looks at the habituation stimulus and may last a fixed maximal duration or may end earlier as soon as the infant looks away from the stimulus. To start a new trial, the renewed presentation of the stimulus is often preceded or accompanied by an attention grabbing stimulus, like a flashing light or a sound effect. Across trials, infants' orientation responses weaken. Habituation trials are repeated during the habituation phase until a criterion is met. Typically, total looking time across three consecutive trials must fall below half the total looking time on the first three habituation trials for the habituation phase to end. In the subsequent test phase new stimuli are presented. Renewed orientation behavior toward such new stimuli is referred to as dishabituation and indicates that habituation is specific to the habituation stimulus. Sometimes, an orientation response continues to be observed when the habituation stimulus is then again presented, a phenomenon referred to as Spencer-Thompson dishabituation (Thompson and Spencer, 1966).

Conceptually, habituation could be viewed as a signature of destabilization where the reduced looking time results from reduced stabilization of visual fixation or, generally, reduced responsiveness (Balkenius, 2000; Sirois and Mareschal, 2002, 2004; Schöner and Thelen, 2006). This is consistent with how habituation manifests itself in preferential looking tasks (Roder et al., 2000) that probe perception in a way that is analogous to how motor decisions are probed in the A not B task. Stimuli at two spatial locations are repeatedly presented to infants. At one location, the stimulus remains the same across repetitions, at the other location it varies and is thus always new to the infant. Orientation is assessed by looking time at either of the two stimuli. Across the first few repetitions, infants tend to look longer at the invariant stimulus, a finding referred to as familiarity preference. After longer exposure, infants tend to look longer at the novel stimulus, a finding referred to as novelty preference.

Familiarity preference may then be viewed as a form of stabilization in which the established spatial orientation resists change to the location of the novel stimulus. Novelty preference would then reflect habituation to the familiar stimulus which destabilizes the orientation response to that stimulus. The pattern of early familiarity and late novelty preference would thus suggest that stabilization predominates early during such repeated stimulation, while destabilization prevails later.

This is how neural dynamic models provide theoretical accounts for both perseverative reaching (Thelen et al., 2001; Dineva and Schöner, 2018) and visual habituation (Schöner and Thelen, 2006; Perone and Spencer, 2013b). Neurons tuned to relevant features are modeled at the population level as neural dynamic fields that span the feature dimensions. Localized activation patterns (or peaks) in these fields represent perceptual or motor states. Activation peaks are induced by external input. Once activation exceeds the threshold of neural transmission, a pattern of recurrent, locally excitatory connectivity within the fields begins to stabilize localized activation peaks. Inhibitory recurrent connectivity, neurophysiologically mediated by a field of inhibitory interneurons, supports selective activation at one field location when multiple locations receive input. Once a peak has been induced, activation in both excitatory and inhibitory populations may be strengthened over time due to a simple learning mechanism, modeled as a memory trace. This accounts for effects across multiple presentations or reaches in these models [and corresponds to the "latent memory trace" in the alternative connectionist model of perseverative reaching (Munakata, 1998)].

In the account for perseverative reaching in the A not B paradigm (Thelen et al., 2001; Dineva and Schöner, 2018), the activation field spans the direction of the infants reaching movements. When a reach to the A location is cued on an A trial, input is provided to the location of the field that corresponds to reaches to that location. Once activation at that field location passes the threshold, a reach to A is predicted. The memory trace of the activation field strengthens activation at that location, making it easier to elicit the same movement again on the next trial. This build-up of a memory trace across trials is responsible for perseveration when the B location is cued on a later B trial. Essentially, the reinforced activation pattern for a reach to the A location competes with activation induced by the cue for a reach to the B location. That induced activation decays over a delay, while the memory trace persists, so that the competition is increasingly biased toward the reach to A for longer delays.

In the account for visual habituation (Schöner and Thelen, 2006), the activation field spans visual features of the stimuli presented to the infants. While the infant is looking at a particular stimulus, localized input is provided to the field, inducing a peak of activation. The model postulates that such a perceptual peak stabilizes fixation of the stimulus. The model accounts for habituation by the build-up of a memory trace in the inhibitory layer of the perceptual field. Across trials, inhibition is strengthened, weakening the perceptual representation, and thus its stabilizing influence on fixation. The modeled infant will tend to look away from the stimulus to which it has habituated. Perone and Spencer (2013b) provide a elaborated neural dynamic account of visual habituation, in which the perceptual activation layer drives a working memory for the percept. As perceptual activation is strengthened by a memory trace, working memory passes a threshold. It is this new working memory for the percept that induces inhibition through its inhibitory layer that accounts for the weakening of the perceptual representation over viewing time and predicts looking away. That neural dynamic account of habituation may be seen as consistent with the Sokolov perspective (Sokolov, 1963) and its modern neural network implementation (Sirois and Mareschal, 2004), in which attention to a stimulus is stabilized while perceptual representations are being built, and destabilized thereafter.

In the neural dynamic models, perseveration in the reaching tasks and habituation in perceptual tasks are both caused by the build-up of activation through memory traces, but in different layers: Perseveration results from strengthened activation in an excitatory layer that drive motor behavior. Habituation results from strengthened activation in an inhibitory layer that weakens motor behavior. A unified account would postulate that, generically, memory traces strengthen activation both in excitatory and inhibitory layers. In such a unified account, familiarity preference in perceptual tasks and perseverative reaching in motor tasks originates from the memory trace in the excitatory layer. Habituation originates from the memory trace in the inhibitory layers. The unified account would be valid if habituation was also observed in motor tasks, so that a particular motor behavior becomes less likely when it is being performed repeatedly. Such motor habituation predicts a form of novelty preference, in which a habituated infant would then prefer to perform a new motor behavior over a familiar motor behavior.

Observations by Marcovitch et al. (2002) and Marcovitch and Zelazo (2006) in the A-not-B paradigm are consistent with this suggestion. These studies looked at how the number of reaches to A matters. In the experimental procedure the toy was hidden at the A location for one, six, or eleven trials before switching to the B location. This led to a U-shaped effect: Infants assigned to the single A trial condition did not perseverate at all. Infants in the traditional 6 A trial condition perseverated. Infants in the 11 A-trial condition were less likely to perseverate. The neural dynamic model of perseveration explains the absence of perseveration in the single A trial condition by the limited experience reaching to A, so that only a weak memory trace has been built. The model does not explain the reduced level of perseveration in the 11 A trial condition. A unified model would account for this reduction by the built up of an inhibitory memory trace that reflects habituation of the A reach.

In this article, we report an experiment that employs the experimental procedure of the habituation paradigm in a movement task. The experimental results provide evidence for habituation of movement generation that is specific to the direction of the movement: When the movement direction changes, we observe dishabituation. Moreover, we find a motor variant of Spencer-Thompson dishabituation. We then introduce a neural dynamic model that unifies previous accounts for habituation (Schöner and Thelen, 2006; Perone and Spencer, 2013b) and perseveration (Thelen et al., 2001; Dineva and Schöner, 2018). We use the model to account for the experimental finding. Finally, we extrapolate the model to a paradigm that involves motor selection in which the model accounts for perseverative reaching in the A-not-B paradigm (Smith et al., 1999) and the reduction of perseveration with increasing experience of an initial choice (Marcovitch et al., 2002).

2. MOTOR HABITUATION EXPERIMENT

The motor habituation experiment mimicked the visual habituation paradigm. A box with a lever was repeatedly presented to toddlers (see **Figure 1**). Depending on how the box was presented, moving the lever entailed vertical or horizontal movements of the hand. Moving the lever lead to the box playing music and was, therefore analogous to fixating a stimulus in the visual habituation paradigm, which leads to visual stimulation. Only one movement direction was possible at a given time, the box's orientation was altered between habituation and test trials to probe for dishabituation and Spencer-Thompson dishabituation. Analogous to the A-not-B paradigm, action was elicited by pushing the box into the reaching space of toddlers and action was terminated by pulling the box out of reach when a trial ends.

2.1. Method

2.1.1. Participants

Thirty eight 12-month-olds (23 boys, 16 girls) and 38 15-month-olds (22 boys, 17 girls) toddlers participated. Twenty one other toddlers were recruited but did not finish the experiment due to fussiness or technical problems. Their data were not included in the analysis. Toddlers of each age group were randomly assigned to two experimental conditions (starting with horizontal/vertical movement), resulting in 19 toddlers in each condition and age group.

2.1.2. Apparatus and Data Acquisition

A lever mounted in the center of a box could be slid along a notch with a maximal range of motion of 11 cm (**Figure 1**). To minimize visual distraction and the influence of perceptual habituation, the box was deliberately made visually boring, painted black with two yellow stripes parallel to the notch indicating the movement direction. We don't expect toddlers to habituate to such boring visual stimuli. Through a Labview data acquisition program, a computer recorded the moments in time when the lever was being moved and its current displacement. Based on the movement data, the computer controlled the speaker in the box, playing a sound file (Vivaldis piccolo concerto in c major) whenever the lever was being moved and turning it off when the movement stopped.

The box was placed on a board whose tilt angle relative to the table on which it was mounted could be adjusted to set the movement direction of the lever to horizontal or to vertical. The board could also be moved by the experimenter along a track closer or further away from the toddler. A semicircular notch cut out on the front of the table enabled the toddler to comfortably sit on a parent's lap facing the table and the box (see **Figure 1**). The parent sat on a rolling chair and positioned the toddler close

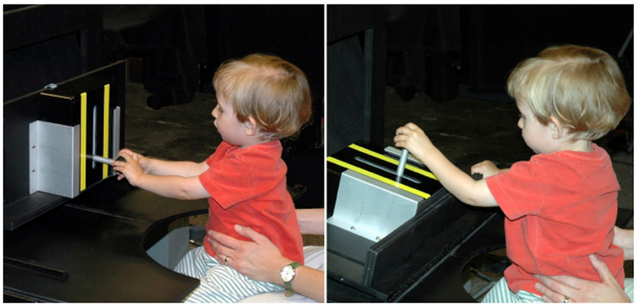


FIGURE 1 | Experimental setting: A box with a lever was mounted on a black board with aluminum braces. The box could be oriented to enable vertical or horizontal movement of the lever. The box and environment were visually nondescript, besides two yellow stripes indicating the movement direction. While the lever was being moved, the box played music.

to the edge of the table. The experimenter sat cross the table from the toddler, hidden by a curtain to reduce distraction.

During the experiment, the displacement of the lever was displayed on a screen in real time and an LED indicated whether the lever was being moved. The current trial number and elapsed time on the current trial were displayed and updated online. The total moving time and a habituation criterion were calculated online and used to control the timing of the experiment. The total moving time on the first three trials and the last three trials were displayed together with their ratio in percent. When the total moving time on the last three habituation trials fell below 50 % of the total moving time on the first three habituation trials, an LED labeled "Reached Criterion" flashed. The experimenter then stopped the habituation phase by withdrawing the box, changed the angle of the box, and began the test trials.

The entire experimental session was videotaped for later review with two video cameras mounted in front of and on the right side of the toddlers, respectively.

2.2. Procedure

In the horizontal condition, toddlers of both age groups were first habituated to the horizontal movement direction, tested with the vertical movement direction on the first two test trials, and then tested again with the horizontal movement direction in two additional test trials. In the vertical condition, the same sequence was run through with horizontal and vertical movement direction swapped. Each trial lasted 15 s. The toddler-controlled habituation criterion determined the end of the habituation phase, when the total moving time on the last three habituation trials fell below 50 % of the total moving time on the first three habituation trials. This way, we apply the classic habituation criterion widely used in visual habituation to a motor task (see Colombo and Mitchell, 1990 for an overview of paradigms/criteria). The test phase started when the habituation criterion was met or the toddler finished 15 habituation trials.

Two experimenters were needed to run the experiment. Experimenter 1 operated the computer and informed experimenter 2 when a trial terminated. Experimenter 2 hid behind the curtain, withdrew and retrieved the box between trials, and changed the tilt angle of the board at the transition from the habituation to the test phase, and from test trial 2 to test trial 3. Experimenter 2 made the inter-trial interval constant through practice, which was about 11.5 ± 1.5 s across all trials, including the transition from one movement direction to the other.

After the parent completed the consent documents and the toddler was comfortable in the lab, parent and toddler sat down in front of the box. The toddler was given a short period of time to get familiar with the box before data collection started. The movement direction of the lever during this warm-up phase was the same as that in the following habituation phase. It followed a strict routine: The parents demonstrated the movement twice, held the toddlers hands on the handle twice, and then encouraged the toddlers to move the lever themselves. After the toddlers moved the lever independently for three times, the box was pulled back and pushed into place again to start the experimental trials.

At the beginning of each trial, the parents drew the toddlers attention to the box and put their hands on the lever if the toddlers did not voluntarily do so. Experimenter 1 started the 15 s trial in the Labview program once the toddlers hands were on the knob of the lever. During a trial, the parents were asked not to interact in any way that would influence or distract the toddlers. However, they were allowed to say encouraging words when the toddlers moved the lever.

After the habituation criterion was met or the maximum of 15 habituation trials was exceeded, the orientation of the box was changed and two test trials started. Toddlers watched the experimenter rotating the box. No warm-up was given for the novel movement direction, the test trials started immediately after the last habituation trial. After two test trials in the new movement direction, the box was changed back to the familiar direction for two additional test trials.

2.3. Data Analysis

2.3.1. Habituation and Dishabituation

Toddlers started and stopped moving the lever several times within a trial. The movement times of those episodes were summed for each trial. The habituation criterion was defined in terms of summed movement time as described above. Only 11 of 76 toddlers did not reach the habituation criterion so that their habituation phase ended after 15 trials. Dishabituation and Spencer-Thompson dishabituation were assessed through *t*-tests that compared the movement times in the test trials with the movement times of the last habituation trial for each age group and habituation condition.

As a second measure, the movement paths of all episodes within a trial were summed. Habituation manifests itself in movement path as well, when the habituation criterion defined on the basis of movement path is satisfied for the movement path in the last habituation trials. Dishabituation and Spencer-Thompson dishabituation were assessed based on movement path for each age group and habituation condition by *t*-tests.

2.3.2. Handedness

Some toddlers switched hands across trials during the experiment. The hand toddlers used in each trial was coded from the video tape as left hand, right hand, or both hands. There were more hand switches in early trials than in late trials. Out of 76 participants, 26 switched hands during the experimental procedure (14 12-month-old, 12 15-month-old).

The decrease in movement time across subsequent trials with and without a hand switch was analyzed for those toddlers who switched hands. Only trials during the habituation phase were considered. The decrease of movement time on early trials (first three habituation trials) was compared to the decrease of movement time on late trials (last three habituation trials) for trials with and without a hand switch and for all age groups and conditions in an ANOVA.

2.4. Results

2.4.1. Habituation and Dishabituation

Figure 2 shows the average movement times on the first three and last three habituation trials and on the test trials for each age group and condition. Average movement times decrease during the habituation phase, satisfying the habituation criterion in all age groups and conditions. When the new movement direction is tested, average movement times increase compared to the last habituation trial. This provides evidence for dishabituation, which is significant, at p-value < 0.05, in all age groups and conditions in the first and second test trial (see orange box in **Figure 2**). When the original movement direction is tested again in test trials three and four, movement time increases sightly compared to the last habituation trial. This provides evidence for Spencer-Thompson dishabituation, which is significant for all age groups and conditions on test trial three. On test trial four, it is significant only for 15-month-old in the vertical habituation condition (see green boxes in **Figure 2**).

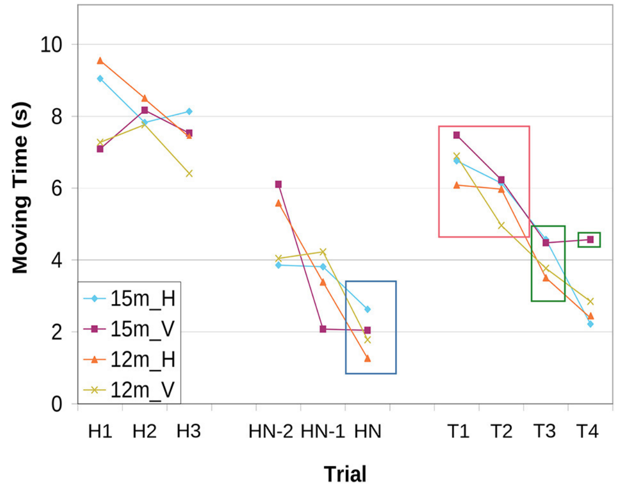


FIGURE 2 | Experimental results: Average movement times for the first (H1-H3) and last three habituation trials (HN-2, HN-1, HN), as well as the test trials (T1-T4). Movement times are averaged across age group (12 or 15 months) and habituation condition, horizontal (H) or vertical (V). Average movement time satisfies the habituation condition in all groups in the last habituation trial, HN, marked by the blue box. The orange box marks significant dishabituation (difference to last habituation trial, HN) to the new movement direction in T1 and T2. Spencer-Thompson dishabituation is significant (difference to HN) in some trials and for some age groups/conditions, marked by the green boxes.

Results based on the second measure of movement, the summed movement path per trial, have the same structure: The habituation criterion is met on the last habituation trial in all groups/conditions. Average movement paths lengthened on test trials one and two. This dishabituation to the new movement direction was significant in all age groups and conditions. In test trials three and four, average movement paths lengthened slightly compared to the last habituation trial. Spencer-Thompson dishabituation was significant in T3 for all age groups and conditions, in T4 only for 15-month-old in the vertical movement direction.

These results provide evidence for habituation to movements. The observed dishabituation shows that habituation is specific to a specific movement direction, suggesting the existence of novelty preference in motor behavior.

2.4.2. Handedness

The only significant main effect on change of movement time across subsequent trials reflected that movement time decreases more strongly in late trials than in early trials. The decrease in movement time did not interact with age or condition, nor does it interact with the presence or absence of a hand switch. **Figure 3** shows the distribution of movement time decreases from one trial to the next when a switch of hand occurred as contrasted to movement time decreases from one trial to the next when no switch of hand occurred. These distributions are shown separately for trials early and late during habituation.

The result suggests that habituation is not specific to the effector used. Such a dependence would predict less decrease or an increase of movement time after a hand switch. This is consistent with ascribing habituation to a level higher than the effector specific movement generator, for example, to a level representing an intention to move the lever. This informs the choice of level of description in the model.

3. NEURAL DYNAMIC MODEL OF MOTOR HABITUATION

3.1. Motor Habituation Model

To account for motor habituation as observed in the reported experiment, we unify previous neural process accounts for perceptual habituation (Schöner and Thelen, 2006; Perone and Spencer, 2013b) and perseverative reaching in the A-not-B paradigm (Thelen et al., 2001; Dineva and Schöner, 2018) that were based on the framework of Dynamic Field Theory (DFT) (Schöner et al., 2016). A two-layer neural dynamic field is defined over movement direction x (see **Figure 4**). At the first layer, an excitatory field, $u(x, t)$, represents the intention to move in a particular direction, x. At the second layer, an inhibitory field, $v(x, t)$, mediates habituation. It receives excitatory input from the intention field, $u(x, t)$, which it in turn inhibits. Activation in

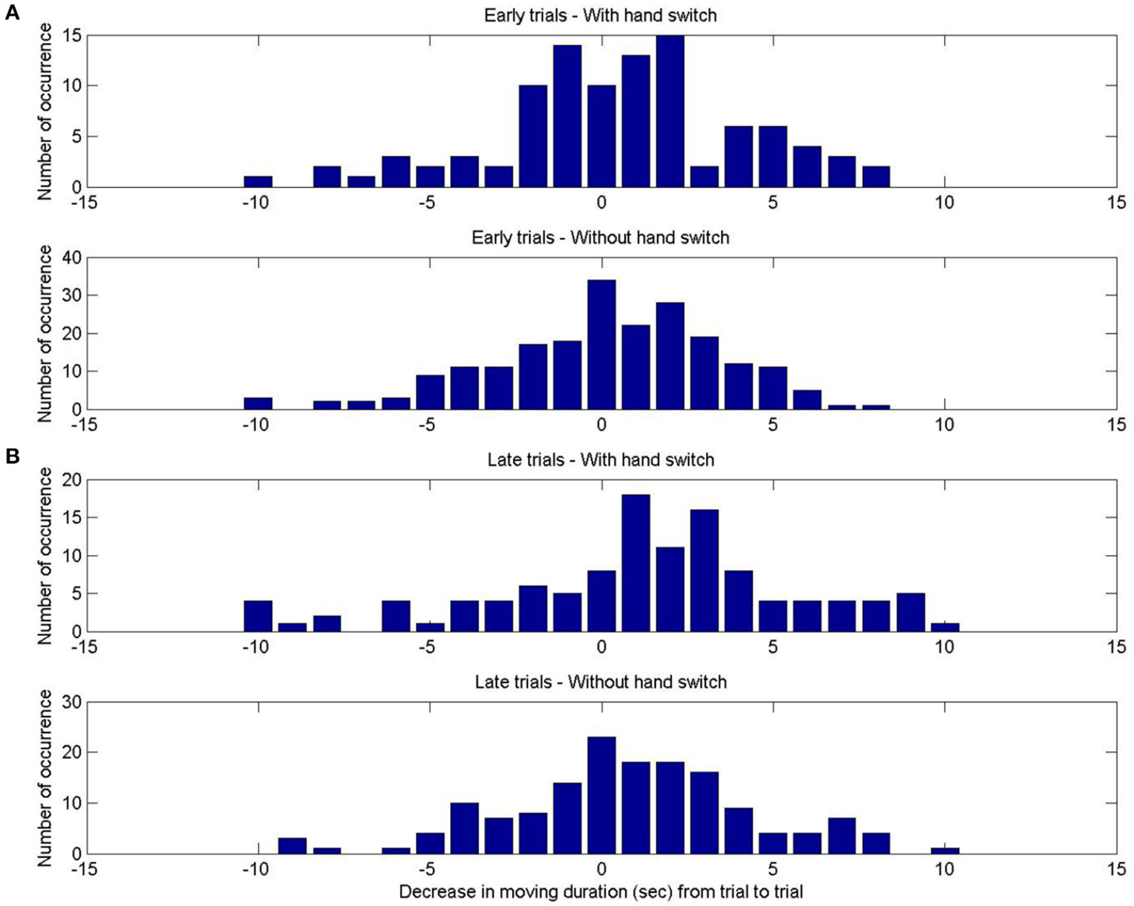


FIGURE 3 | Experimental results: Distribution of decrease in movement time between two trials. **(A)** In the first three habituation trials, when a hand switch occurred between trials (top) and without hand switch (bottom). **(B)** In the last three habituation trials, when a hand switch occurred between trials (top) and without hand switch (bottom).

both fields evolves continuously in time as described by neural dynamics (described in mathematical detail below, see Equation 1 in section 3.1.1).

The intention field evolves under the influence of a variety of inputs, $s(x,t)$, that reflect perceptual information (see below). Recurrent connectivity within the fields contributes more strongly than such external inputs, however. Local excitatory connectivity within the intention field stabilizes localized patterns of activation against decay. Input from the inhibitory field stabilizes peaks against diffusive spread, but may also weaken activation patterns in the intention field. Only field locations that are sufficiently activated engage in neural interaction, as modeled by a sigmoid threshold function that makes the neural dynamics nonlinear (see Equation 4).

Without input, activation in both fields is at a negative resting level. When input pushes activation at some field location through the threshold of the sigmoid, the sub-threshold pattern of activation becomes unstable. Driven by local excitatory interaction, activation evolves to a supra-threshold stable state, that is, a localized peak of activation. In the excitatory field, this represents the intention to move the hand in a particular direction that is encoded by the location of the peak along the field dimension[1].

Various perceptual inputs to the intention field model the experimental procedure. Task input, s_T, represents that a box affording a particular movement direction is within reach. The trials and inter-trial intervals are modeled by varying task input in time. Reward input, s_R, models the strengthening of an active movement intention when the rewarding outcome, the music, is perceived. In the simulations, this input is only present while a supra-threshold peak exists that would induce lever movement in the (unmodeled) motor system. A third input models the parent's action of drawing attention to the box and encouraging the child to move the lever. This attention input, s_A, is applied while task input is provided to the intention field but no supra-threshold peak has yet formed.

[1] How such intentional activation may actually drive movement generation down to activating muscles is not modeled here. See Schöner et al. (2019) for a sketch of such a more complete DFT model of the generation of reaching movements.

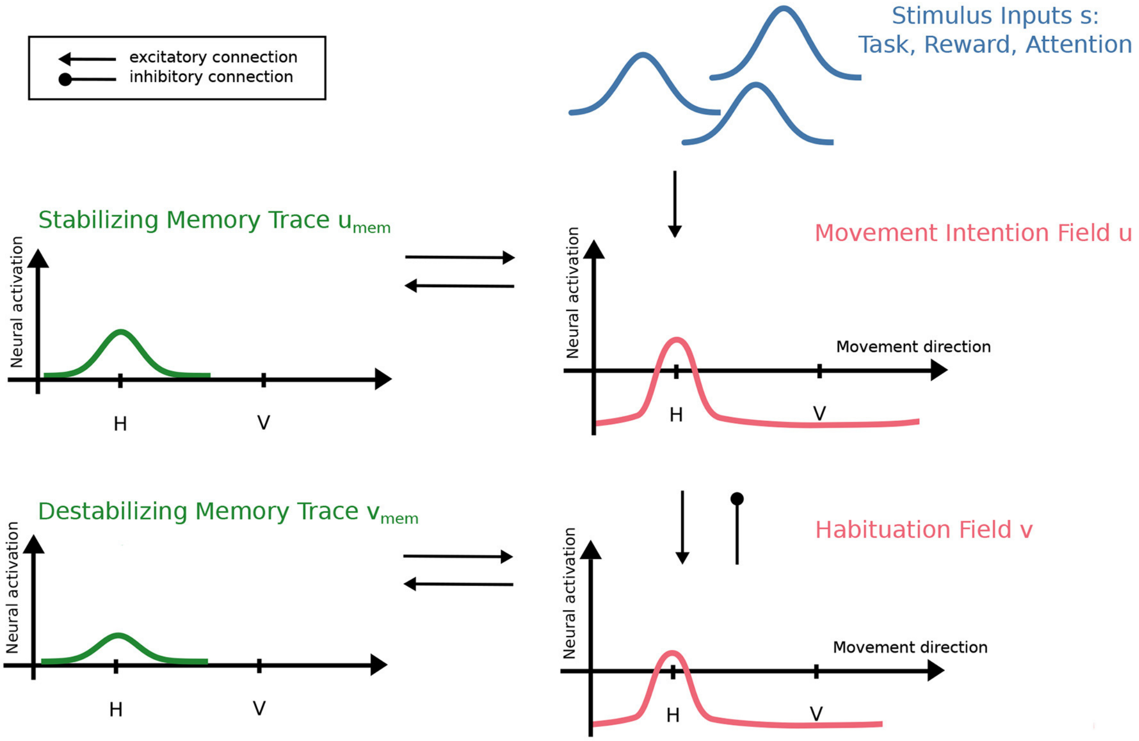


FIGURE 4 | DFT model of motor habituation. The intention field, *u*, receives stimulus inputs, *s*, that models the visual perception of the box and lever, the perception of rewarding outcomes, or stimulation by a parent. The intention field provides input to the habituation field, *v*, which in turn inhibits the intention field. Both fields are defined over the movement direction, *x*, sampled at horizontal (H) and vertical (V) movement directions in the experiment. A supra-threshold peak at a location, *x*, in the intention field indicates that a movement in that direction is being generated. Memory traces reflect the recent history of supra-threshold activation in both fields and provide input back to the fields. They facilitate peak formation and, thus, account for the stabilization and destabilization of movement intentions.

It may be strong enough to push the intention field through the detection instability.

A peak in the intention field decays when the supra-threshold state becomes unstable so that activation falls back to a sub-threshold state. This happens in the reverse detection instability at lower levels of input than the detection instability. The decay of a peak reflects the decision to stop moving the lever. This happens when the task input is removed at the end of a trial or when inhibitory input from the habituation field becomes sufficiently strong.

Habituation (and perseveration) reflect the history of activation. The model represents that history through memory traces of both activation layers, *u* and *v*, of the model. In DFT, dynamic memory traces model a simple form of learning (akin to the dynamics of the bias inputs in connectionist networks). The memory trace builds on a slower time scale at locations with supra-threshold activation (see Equation 5) and decays if those locations fall below threshold while supra-threshold activation is present anywhere else in a field. Without supra-threshold activation in a field, the memory trace remains constant. Memory traces act like a locally enhanced resting level, preshaping the activation patterns in the field and facilitating peak formation at these locations. The memory trace, u_{mem}, of the intention field thus accounts for the stabilization of movement intentions. The memory trace v_{mem}, of the habituation field accounts for the destablization of movement intentions by enhancing inhibition.

When there are localized inputs at multiple field locations, only one peak may form in the intention field due to inhibitory input from the habituation field. The motor habituation experiment does not probe such selection decision as only a single movement direction is afforded at any moment in time. We will examine situations involving selection in the model, to connect the account to models of motor decision. Memory traces in excitatory fields of such models have previously been used to account for pre-trial effects (Erlhagen and Schöner, 2002; Dineva and Schöner, 2018), and perseveration (Thelen et al., 2001).

In DFT models of visual habituation, activation in the excitatory perception field is defined over features of the visual percept (Schöner and Thelen, 2006; Perone and Spencer, 2013b). This perceptual activation is assumed to stabilize fixation. Reduced activation due to the build-up of inhibition then promotes looking away, a signature of habituation (Schöner and Thelen, 2006; Perone and Ambrose, 2016), and preferential looking (Goldberg and Schöner, 2007; Perone and Spencer, 2013a,b).

3.1.1. Mathematical Formulation

The evolution of the intention and habituation fields is modeled by this neural dynamics:

$$\begin{aligned}\tau_{\mathrm{u}}\dot{u}(x,t) &= -u(x,t) + h_{\mathrm{u}} + s(x,t) + \int k_{\mathrm{uu}}(x-x')g(u(x',t))\,\mathrm{d}x' \\ &- \int k_{\mathrm{uv}}(x-x')g(v(x',t))\,\mathrm{d}x' \\ &+ \int k_{\mathrm{uu_{mem}}}(x-x')g(u_{\mathrm{mem}}(x',t))\,\mathrm{d}x' + \tau_{\mathrm{u}} q \xi_{\mathrm{u}}(x,t), \quad (1)\\ \tau_v \dot{v} &= -v(x,t) + h_{\mathrm{v}} + \int k_{\mathrm{vu}}(x-x')g(u(x',t))\,\mathrm{d}x' \\ &+ \int k_{\mathrm{vv_{mem}}}(x-x')g(v_{\mathrm{mem}}(x',t))\,\mathrm{d}x' + \tau_{\mathrm{v}} q \xi_{\mathrm{v}}(x,t).\end{aligned}$$

Independent Gaussian white noise, $\xi_i(x,t)$, with strength q is applied to all field locations. The time scales, τ_i, determine how fast activation in the fields evolves. Without inputs, activation in the fields is at the negative resting level $h_i < 0$. The input, $s(t,x)$, sums over the three sources of stimulation and is applied to the intention field u during the experimental procedure. Stimulus components, $s_k(x)$, are modeled as Gaussian functions:

$$s_k(x) = \frac{a_k}{\sqrt{2\pi}\sigma_{\mathrm{exc}}} \exp\left\{-\frac{(x-x_0)^2}{2\sigma_{\mathrm{exc}}^2}\right\}, \quad (2)$$

with width σ_{exc} and amplitude a_k. The index k = T,R,A corresponds to the Task, Reward, or Attention input. The Gaussian functions are centered on $x_0 = H$ or $x_0 = V$ for a horizontal or vertical movement direction, respectively.

Lateral interactions within and between the fields are determined by interaction kernels, k_{ij}

$$k_{ij}(x-x') = \frac{c_{ij}}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x-x')^2}{2\sigma_{ij}^2}\right\} + c_{ij,\mathrm{glob}}, \quad (3)$$

where the first index corresponds to the target field and the second to the source field of the projection. The Gaussian part models local interaction within a field ($i = j$) or coupling to other fields (from field j to field i) with width σ_{ij} and strength c_{ij}. Global interaction is determined by the constant $c_{ij,\mathrm{glob}}$ which is applied to all field locations.

Only field locations that have sufficient levels of activation engage in lateral interaction. The output of a field u is determined by a sigmoid function with threshold at zero, whose steepness is given by β:

$$g(u) = \frac{1}{1+\exp(-\beta u)}. \quad (4)$$

The memory trace of the intention field grows with the time scale τ_{build} more slowly than the fields:

$$\begin{aligned}\dot{u}_{\mathrm{mem}}(x,t) =& \tau_{\mathrm{build}}^{-1}\left[-u_{\mathrm{mem}}(x,t) + g(u(x,t))\right] g(u(x,t)) \\ &-\tau_{\mathrm{decay}}^{-1} u_{\mathrm{mem}}(x,t)\left[1 - g(u(x,t))\right],\end{aligned} \quad (5)$$

as long as there is supra-threshold activation at any location in the corresponding field. Otherwise, the memory trace remains constant ($\dot{u}_{\mathrm{mem}} = 0$). Activation in the memory trace thus decays competitively only when there is supra-threshold activation at other field locations. In general, the time scale for decay, τ_{decay}, is slower than for building the memory trace. The dynamics of the memory trace, v_{mem}, of the habituation field is described by the same dynamics, although the time scales may differ.

3.1.2. Constraints on Model Parameters

The experimental procedure, observations during the experiment, and qualitative assumptions about the results provide constraints for setting many of the parameter values of the model:

(1) Toddlers moved the lever only after the warm-up phase during which they were encouraged by their parent. We assume this to be a critical part of the procedure that enabled the toddlers to associate the lever moving action with the rewarding outcome, the music. We expect that they would not be interested to move the lever without the music. The amplitude of the task input, s_{T}, is chosen, therefore, such that task input alone is not sufficient to elicit a supra-threshold peak in the intention field. Only task input in combination with input from the stabilizing memory trace or the attention input induces supra-threshold activation in the intention field.

(2) Since toddlers do not try to move the lever while the box is out of reach, input from the stabilizing memory trace, u_{mem}, to the intention field alone is assumed to be insufficient to induce a peak. This constrains the coupling strength, $c_{\mathrm{uu_{mem}}}$, to be less than the absolute value of the intention field's resting level $|h_{\mathrm{u}}|$. A combination of at least two of the three sources of inputs, task input, attention input, and input from the excitatory memory trace is assumed necessary to induce a detection instability in the intention field.

Since the rewarding input is only applied when there already is supra-threshold activation in the intention field, it does not play a role in inducing a detection instability. However, it further stabilizes the decision when the attention input is removed. This models that toddlers who were encouraged to move the lever at the beginning of a trial kept moving when perceiving the rewarding music without a need for continued stimulation from their parent.

(3) Typically, after a few trials toddlers stopped moving the lever even while the box was within reach. The coupling strength, c_{uv}, from the habituation to the intention field is thus chosen such that the inhibitory input to the intention field becomes larger than the sum of task input and input from the memory trace, u_{mem}. This makes it possible that a supra-threshold peak in the intention field can be destabilized by inhibition from the habituation field.

At the beginning of a trial, the parent encourages his or her child to move the lever. The coupling strength, c_{uv}, is thus assumed to be smaller than the attention input combined with the task input and input from the stabilizing memory trace so that the attention input may elicit a peak in the intention field despite strong inhibition from the habituation field.

(4) Since there is no self-excitation within the inhibitory layer, the coupling strength, c_{vu}, must be strong enough for

the intention field to cause supra-threshold activation in the habituation field.

(5) To model Spencer-Thompson dishabituation, the destabilizing memory trace of the habituation field, v_{mem}, must decay faster than that of the intention field. Thus, after a new movement was performed there is less inhibition at the field location to which the model was habituated.

Supra-threshold activation at another location of the habituation field is necessary for the memory trace, v_{mem}, to decay at an initial location. To obtain Spencer-Thompson dishabituation, a stimulus that is sufficiently different from the initial stimulus must thus be presented after habituation. This constrains the metric overlap between field locations and the respective widths of projection kernels.

(6) The stabilizing memory trace of the intention field, u_{mem}, must grow faster than the destabilizing memory trace of the habituation field, so that it is predominant in early trials. The coupling strength from the habituation field to the intention, c_{uv}, must be stronger than its coupling to the stabilizing memory trace, so that habituation prevails in later trials. This cannot be deduced directly from the motor habituation data, but is consistent with the pattern of early familiarity and a late novelty preference found across a variety of selective tasks.

(7) The experimental results show that the response to the new movement direction is stronger on the first test trial than for the old movement direction on the last habituation trial, but typically not as strong as on the first habituation trials. This points to the existence of global component of habituation across movement directions. Thus, the projection kernel from the memory trace v_{mem} to the habituation field is assumed broader than the projection kernel from u_{mem} to the intention field, including a global (=constant) component.

Table 1 provides an overview of the set parameter values.

TABLE 1 | Parameter values of the habituation model.

Parameter	Value [a.u.]	Meaning/Constraints
β	6	Steepness of sigmoid function
τ_{u}	40	Time scale of u
h_{u}	−1.2	Resting level, $\lvert h_{\text{u}} \rvert \geq s_{\text{T}}$
c_{uu}	1.2	Local excitation in u, stabilizes peak decisions in u
σ_{uu}	2.5	Width of excitatory kernel, $\sigma_{\text{ij}} \ll$ field size for distinct peaks
$c_{\text{uu}_{\text{mem}}}$	0.8	Local input from memory trace, facilitates peak formation at familiar locations
$\sigma_{\text{uu}_{\text{mem}}}$	2.5	Width of excitatory kernel
$c_{\text{uu}_{\text{mem}},\text{glob}}$	0.2	Global input from memory trace $c_{\text{uu}_{\text{mem}}} + c_{\text{uu}_{\text{mem}},\text{glob}} \leq \lvert h_{\text{u}} \rvert \rightarrow$ no spontaneous movement without stimulus inputs
c_{uv}	−1.8	Local inhibition from v, leads to habituation
σ_{uv}	5	Width of inhibitory kernel, broader than excitatory kernel
$c_{\text{uv,glob}}$	−0.4	Global inhibition from v, for habituation and selection decisions $\lvert c_{\text{uv}} \rvert + \lvert c_{\text{uv,glob}} \rvert \geq s_{\text{T}} + c_{\text{uu}_{\text{mem}}} + c_{\text{uu}_{\text{mem}},\text{glob}}$ for "full" habituation
τ_{v}	2	Time scale of v, fast inhibition for global inhibition $\tau_{\text{v}} \ll \tau_{\text{u}}$
h_{v}	−1.2	Resting level, $\lvert h_{\text{v}} \rvert \leq c_{\text{vu}}$ so that u drives supra-threshold activation in v
$c_{\text{vv,glob}}$	−0.1	Global inhibition in v
c_{vu}	2.5	Local excitation from u, drives activation in v
σ_{vu}	2.5	Excitatory kernel width
$c_{\text{vv}_{\text{mem}}}$	3	Local excitation from v_{mem}, modulates strength of habituation
$\sigma_{\text{vv}_{\text{mem}}}$	2.5	Excitatory kernel width
$c_{\text{vv}_{\text{mem}},\text{glob}}$	0.35	Global excitation from v_{mem}, modulates strength of habituation and Spencer-Thompson dishabituation
$\tau_{\text{u}_{\text{mem}},\text{build}}$	200	Building time scale of stabilizing memory trace $\tau_{\text{u}_{\text{mem}},\text{build}} \gg \tau_{\text{u}}$
$\tau_{\text{u}_{\text{mem}},\text{decay}}$	2,000	Decaying time scale of stabilizing memory trace
$\tau_{\text{v}_{\text{mem}},\text{build}}$	600	Building time scale of destabilizing memory trace $\tau_{\text{v}_{\text{mem}},\text{build}} \gg \tau_{\text{u}_{\text{mem}},\text{build}}$ for familiarity preference
$\tau_{\text{v}_{\text{mem}},\text{decay}}$	1,000	Decaying time scale of destabilizing memory trace $\tau_{\text{v}_{\text{mem}},\text{decay}} \leq \tau_{\text{u}_{\text{mem}},\text{decay}}$ for Spencer-Thompson dishabituation
s_{T}	1.0	Task input
s_{R}	1.0	Reward input
s_{A}	1.5	Attention input

Parameters not shown were set to zero in the simulation. The third column addresses the meaning of parameters in the model or constraints we defined for a parameter. For a detailed analysis of parameter constraints see section 3.1.2.

3.2. Simulations

For numerical simulation, the model was implemented in MATLAB using the toolbox COSIVINA for dynamic field architectures[2]. The simulation emulated the procedure of the motor habituation experiment. In the habituation phase, the Gaussian task input, s_{T}, is repeatedly applied to the intention field at location representing horizontal movement, indicating both that the box is in reach and affords a horizontal movement direction. Attention input is added to the same field location when activation does not reach supra-threshold activation within 5 s. On the first trial it is not possible to induce a peak in the intention field because there is no input yet from the stabilizing memory trace u_{mem}. This is when attention input is applied simultaneously with task input, pushing the intention field through the detection instability. This accounts for the warm-up phase of the experiment.

Once a peak forms in the intention field or the attention input is applied, the trial starts. Task input is maintained for another 15 s from that moment on. In the experiment a trial started as soon as the toddlers had their hands on the lever and lasted from then on 15 s. The reward input is added as soon as activation in the intention field reaches the threshold. Any attention input is then removed. At the end of the trial all stimulus inputs are removed for an inter-trial period of 12 s before the task input is applied again at the same field location. With activation reaching the threshold or the attention input added, a new trial begins.

On each trial, the number of time steps at which supra-threshold activation is observed in the intention field are accumulated as a measure for movement time. The simulated movement time is based on the intention to move alone

[2] see www.dynamicfieldtheory.org for access to the sources.

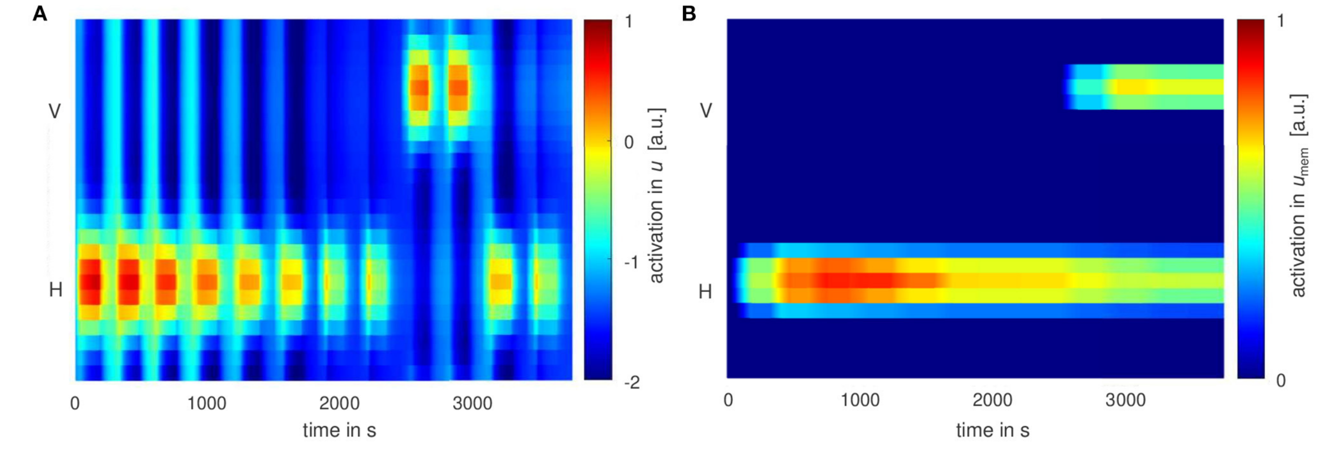


FIGURE 5 | (A) Evolution of activation in the intention field u. Supra-threshold activation (orange-red color) is caused by stimulus inputs applied to the respective field locations repeatedly. **(B)** Evolution of the corresponding memory trace u_{mem}. The memory trace grows at supra-threshold field locations while it decays at all other locations.

(neglecting to model actual movement generation). As in experiment, the habituation phase ends when the habituation criterion is met, that is, the movement time of the just previous three trials is less than 50% of the movement time of the first three trials, or after a maximum of 15 trials.

In the subsequent test phase, the task input is applied twice at the new, vertical field location, modeling that the box is in reach but was rotated. Then, task input is again applied twice at the original field location to probe Spencer-Thompson dishabituation. The trial and inter-trial periods as well as the conditions for applying the attention and reward input remain unchanged.

3.3. Results

3.3.1. Simulation Results in the Habituation Paradigm

Figure 5 shows an exemplary time course of activation in the movement intention field u as well as its memory trace u_{mem}. Time courses of different simulation runs vary due to noise in the fields. Field parameters remained the same in all simulations. Once task and attention input (not shown in **Figure 5**) are applied at field locations representing a horizontal or vertical movement direction, activation at those locations becomes supra-threshold (orange-red color in **Figure 5A**). Supra-threshold activation in u corresponds to the intention to move the lever. Between trials, when no inputs are applied to u, activation in the movement intention field remains sub-threshold (green-blue color). This corresponds to the observation that toddlers did not try to move the lever when the box was out of reach.

During the habituation phase, activation at horizontal field location becomes supra-threshold repeatedly. Over time, amplitude and time-duration of such peaks decrease (areas of red-orange color are narrower than in the first trials in **Figure 5A**) because of increasing inhibition from the habituation field (not shown in **Figure 5**). We assume that supra-threshold activation in the movement intention field u leads to movement generation, that is moving the lever in horizontal direction. Thus, the time of supra-threshold activation in u correlates with movement time measured in the motor habituation experiment. We also expect the amplitude of supra-threshold activation to modify movement generation. It was observed that toddlers moved the lever in several moving episodes rather than moving it continuously during a trial. The amplitude may modify the length of such episodes or the moving speed during an episode.

After the habituation criterion was met, input is applied to vertical field location. Amplitude and time of supra-threshold activation in u are reinstated (red-orange area is broader than in the previous trials) because inhibition from the habituation field is not as strong as at horizontal field location yet. This models dishabituation to a new movement direction. In the last two trials, horizontal task input is applied again and the intention field again becomes supra-threshold at horizontal field location. Inhibition from the habituation field is decreased compared to the last habituation trial which leads to increased movement time (red-orange area is broader than in the last habituation trial). In the experiment this is observed as Spencer-Thompson dishabituation.

Figure 5B shows how the stabilizing memory trace u_{mem} grows at locations of supra-threshold activation in the movement intention field u. The memory trace grows and decays slower than the field. When u becomes supra-threshold at the new, vertical field location, activation in u_{mem} grows at vertical field locations as well, while decaying at horizontal location. Input from u_{mem} to the movement intention field u facilitates peak formation in u. Activation in the habituation field v is driven by input from the movement intention field u and has a similar pattern as shown in **Figure 5A**. Its corresponding memory trace v_{mem} builds slower than the memory trace u_{mem} but decays faster.

For a detailed analysis of the model, **Figure 6** shows a cut through the movement intention field at horizontal (top) and vertical (bottom) field locations as well as activation in the habituation field, stimulus inputs and corresponding memory

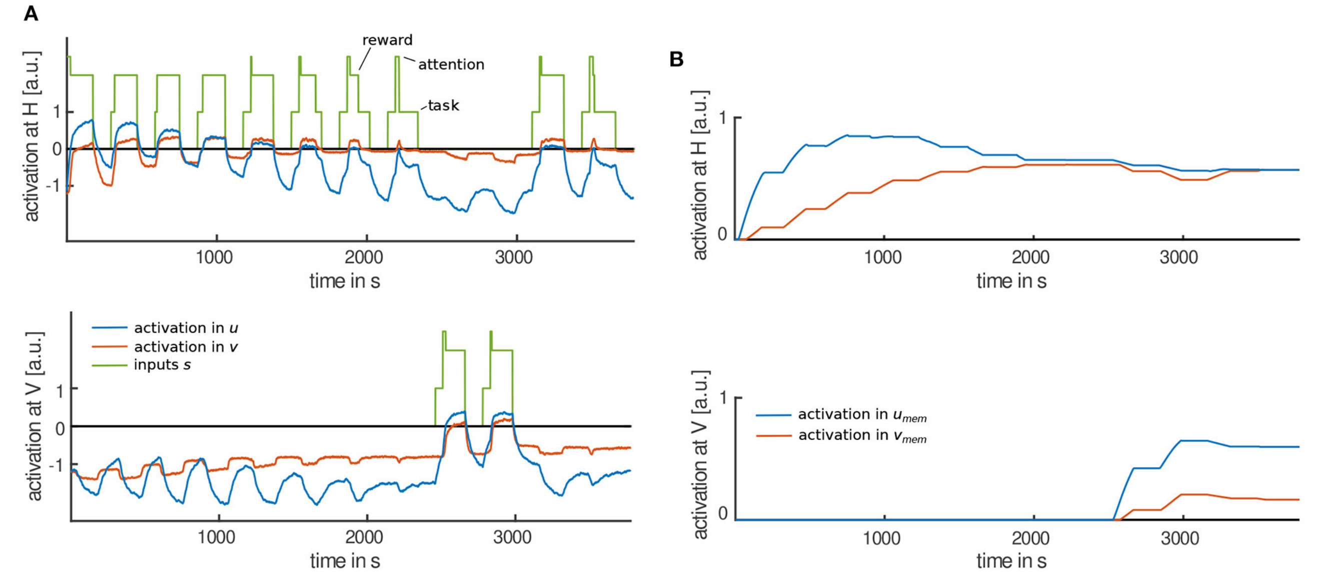


FIGURE 6 | (A) Exemplary time courses of the movement intention u (blue), the habituation field v (red) and stimulus inputs applied (green) at the horizontal (top) and vertical (bottom) field location. **(B)** Corresponding time courses of the memory trace, u_{mem}, of the intention field (blue) and the memory trace, v_{mem}, of the habituation field (red) at horizontal (top) and vertical (bottom) field location.

traces (b). In the first habituation trial, task and attention input are applied to the movement intention field at horizontal movement direction (see **Figure 6A**, top). Once activation in u pierces the threshold of zero, attention input is omitted and reward input is applied. Activation stays above the threshold until task and reward input are removed at the end of the trial. Supra-threshold activation in the movement intention field drives growth of its memory trace (**Figure 6B**, top). This stabilizing memory trace provides input back to the movement intention field so that in the following trials it goes through the detection instability faster and without attention input being applied (trials 2–4 in **Figure 6A**, top). This predicts that toddlers would move the lever spontaneously on these trials.

When the movement intention field becomes supra-threshold input is passed to the inhibitory layer, the habituation field v. Responses in the habituation field are delayed compared to activation in the intention field because it is driven by the intention field only once activation there reaches threshold. Supra-threshold activation in the habituation field drives its memory trace v_{mem}. This destabilizing memory trace v_{mem} provides input back to the habituation field and facilitates peak formation in the following trials, leading to a stronger inhibition of the movement intention field. So, levels of activation in the movement intention field decrease over trials. As a result, activation in the movement intention field does not go through the detection instability, when task input is provided in trials 5–8. Therefore, attention input is applied again. This predicts that toddlers would not move the lever spontaneously on these trials. As a results, parents would need to draw toddlers' attention to the lever.

With increasing inhibition from the habituation field, activation in the intention field is pushed below the threshold even before the trial ends (trials 6–8 in **Figure 6A**). This reproduces the observation in the experiment that toddlers stopped moving the lever although the box was still within reach. The reverse detection instability induced in the intention field is amplified by the removal of the reward input once activation falls below the threshold, making it less likely that the intention field goes through the detection instability a second time. The reward input may also amplify a detection instability, as it is applied once activation in the intention field reaches the threshold, which leads to even higher levels of activation in the intention field. When activation in the intention field goes through the reverse detection instability before a trial ends, movement time decreases. The habituation phase continues until the habituation criterion is met. In the simulation run shown in **Figure 6**, the criterion is met in the eighth trial.

In the first test phase, task input is applied at field locations representing vertical movement direction (see **Figure 6A**, bottom). Again, attention input is needed to push the intention field through the detection instability at the new field location as there is insufficient input yet from the stabilizing memory trace (trials 9 and 10, in **Figure 6B**). Here, the model lacks knowledge that toddlers might actually have about the box playing music even when in a new orientation. However, movement time is reinstated as soon as a peak forms in the intention field and remains until task input is removed. This is how the model accounts for dishabituation to a new movement direction.

In the second test phase, the task input is applied again at horizontal field location of the intention field (see **Figure 6A**, top). Once activation goes through the detection instability, with the help of attention input, it remains supra-threshold for a longer time period than in the last habituation trial. This is because the destabilizing memory trace of the habituation field decays faster than the stabilizing memory trace during the first test trials while task input was applied at the competing field location. Thus, there is less inhibition from the habituation field compared to the last habituation trial, while the impact of the stabilizing memory trace is about the same (see **Figure 6B**, top). In the last test trial, activation in the destabilizing memory trace has grown again and inhibition from the habituation field is strong enough to push activation in *u* through the reverse detection instability before the trial ends. The model thus accounts for Spencer-Thompson dishabituation in the third but not in the fourth test trial.

Figure 7 shows movement times from the model, averaged across 50 simulations runs (analogously to the experimental movement times in **Figure 2**). Because time courses of activation and thus movement times fluctuate across trials, the habituation criterion is met at different trial numbers in different simulation runs. In the first trials, movement time is saturated since activation in the movement intention field remains supra-threshold as long as the 15 s trial lasts.

The model reproduces the reduction of average movement time on the last three trials of the habituation phase over to the average movement time in the first three trials. On the subsequent two test trials, the average movement time is reinstated, a signature of dishabituation. In the second test phase, average movement times are increased in the third test trial (T3) compared to the last habituation trial, a signature of Spencer-Thompson dishabituation. The model shows no significant Spencer-Thompson dishabituation in the fourth test trial.

As stable states, supra-threshold peaks in neural dynamic fields resist noise. Noise may have a strong effect on the system's state near an instability, however. In the model, reward input amplifies small fluctuations when the system is close to the (reverse) detection instability as noise drives activation to positive (or negative) levels. Due to the memory traces, the history of supra-threshold activation has a direct impact on the future time course of activation, leading to variance across simulation runs. [Analogous observations were reported in Perone and Spencer (2013b) in a model of preferential looking.] **Figure 7** reflects this fact through the increase of the standard deviation of movement time increases over trials. In the first two test trials (T1, T2) standard deviation is decreased because activation in memory traces at horizontal field location affect activation at vertical field location only marginally. When task input is again provided at horizontal field location in test trials three and four, standard deviation increases.

3.3.2. Discussion of the Habituation Results

The model simulations are qualitatively in agreement with the experimental data. The model accounts for habituation to a familiar movement direction by a reduced time of movement intention and for dishabituation to a new movement by restoring of movement time. In the third test trial the model also captures Spencer-Thompson dishabituation. We did not try to push quantitative fits beyond what is shown in the figures. The experiment provides evidence for habituation to a movement based on both movement time and movement path. The model operates at the level of movement intentions, so that movement time is accounted for as the time periods during which movements could be generated. Quantitative fits of movement time and path length would need to take processes underlying the actual generation of motor commands into account. These may contribute delays that by themselves depend on the level of activation at the intention level. So, while we expect movement intention to correlate with movement times observed in the experiment, an exact match is not expected. For instance, the modeled movement time is saturated in the first habituation trials (see **Figure 7**) corresponding to the intention or willingness to move throughout the whole trial. In the experiment, movement episodes rather than continuous movements throughout the trial were observed and we expect the actual moving time to be less than the modeled time. **Figure 6A** shows that the amplitude of activation is decreasing in the first habituation trials, probably affecting movement generation and leading to shorter movement episodes. Similarly, the variance induced at the level of movement intention is not necessarily comparable to variance observed at the level of actual movement generation. Moreover, different sources of variation beyond random stochastic perturbations may contribute to experimental assessments of variance, including individual differences (best modeled by differing parameter values), different age groups and variance at the level of sensory inputs.

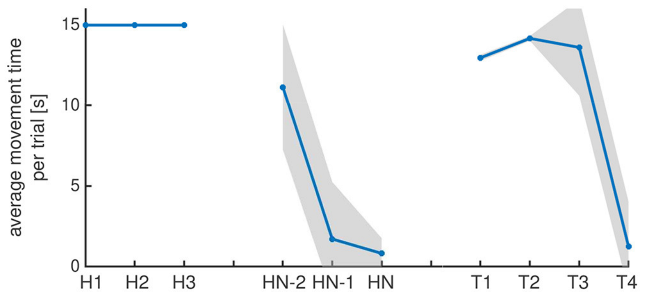


FIGURE 7 | Movement times averaged across simulation runs that were aligned as in the experimental analysis: The first and the last three habituation trials, the test trials in vertical movement direction (T1 and T2) and the test trials in the original movement direction (T3 and T4). The standard deviation across simulation runs due to noise in the fields is shown in gray.

3.3.3. Testing the Effect of Outcome

We expect that toddlers stop moving the lever when the rewarding outcome is suppressed, for example, by no longer playing the music. This was not tested in the experiment presented, but probed in the model by setting parameter values such that task input alone was not sufficient to cause a peak in the intention field. To test how the model behaves when the reward input is omitted in later trials, after stabilizing and destabilizing memory traces have already been built, we modified the simulation procedure. With all parameters of the

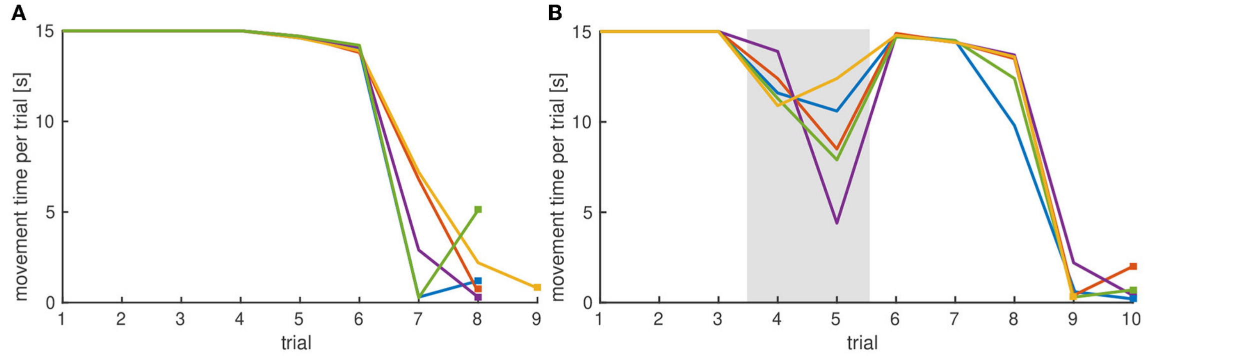


FIGURE 8 | (A) Evolution of movement time over trials for five simulation runs with reward input applied in all trials. The habituation criterion is met at different trials for each run, indicated by the square marker. **(B)** Evolution of movement time over trials for five simulations for which reward input was omitted on trials 4 and 5, marked by the gray area. The habituation criterion is met at later trials, marked by a square.

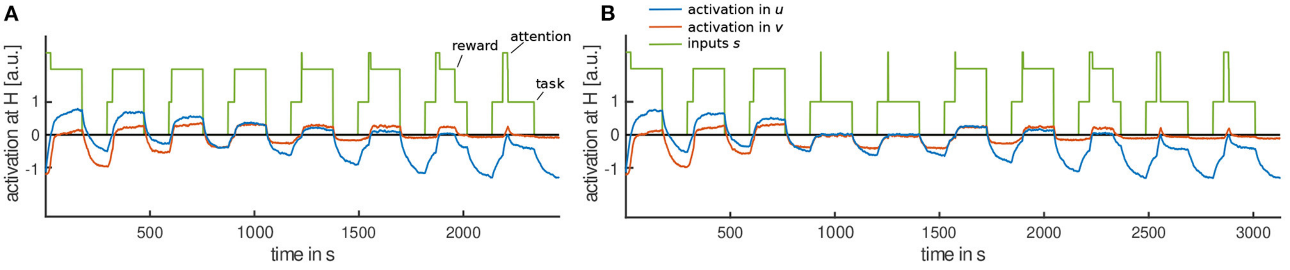


FIGURE 9 | Exemplary time courses of activation until habituation criterion is met: **(A)** With reward input applied on all trials, the habituation criterion is met on the eighth trial. **(B)** When reward input is omitted in trials four to six, the habituation criterion is met on the tenth trial.

model unchanged, the procedure was altered by switching off reward inputs in trials four and five. **Figure 8B** shows that movement time decreased on those trials. This is because lower levels of activation are more easily inhibited by the habituation field. **Figure 9** compares the time courses of activation until the habituation criterion is met with reward input applied in all trials (a) and reward input omitted in trials four and five (b).

A more interesting question might be whether not receiving a rewarding outcome affects the process of habituation. We predict that trials without reward input do not contribute or contribute less to habituation than trials with a rewarding outcome. Habituation criterion would then be met at later trials. Model simulations support this idea: When reward input applied in all trials, the criterion is met after 7.9 (±0.3) trials averaged over 50 simulations. When reward input is omitted on trials four and five, the criterion is met in the 10.0 (±0.2) trial on average.

The model predicts that movement time is decreased in trials without a rewarding outcome. Due to less activation in the movement intention field those trials do not contribute or contribute less to habituation. Movement times are reinstated when the rewarding outcome is perceived again, which "resets" the process of habituation and, thus, the habituation criterion is met in later trials. **Figure 9B** shows that activation in the sixth trial is increased compared to activation when the reward input was applied in all trials (**Figure 9A**).

3.3.4. Simulation of a Selection Task

In selection tasks, a transition from familiarity preference in early trials to a novelty preference in later trials is often observed. In the A-not-B paradigm, perseverative reaching could be viewed as a form of familiarity preference. The findings by Marcovitch et al. (2002) and Marcovitch and Zelazo (2006) show that with more experience of reaching to the A location, infants are less likely to perseverate. This could be viewed as a signature of habituation and a form of novelty preference.

Our experiment did not probe action selection. In the model, we may simulate action selection by simultaneously providing input at two field locations. This simulations can then be compared to the perseverative reaching paradigm. Task input is repeatedly applied to one field location, with a trial duration of 15 s and an inter-trial period of 12 s. In a second phase, an additional input over a second field location is added which competes with the continued input at the first location. This second phase occurs either early or late during habituation to the stimulus at the initial location.

Attention input is only applied in the first trial as it would bias the selection decision to one of the two movement directions. The familiar task input is applied at $x = H$, the additional input is applied at $x = V$. The amplitude of the novel task input, $s_T(V)$,

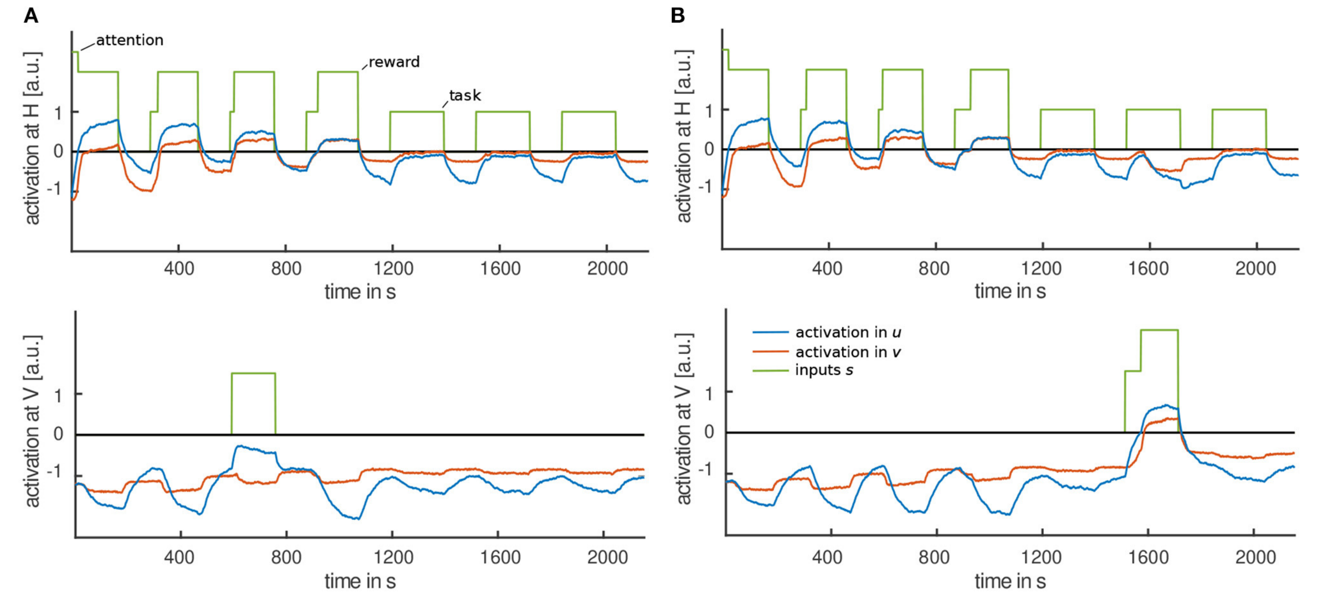


FIGURE 10 | (A) Time courses of activation at the horizontal (top) and vertical (bottom) field locations with an additional stimulus applied at $x = V$ on trial three. The stimulus at the familiar location (top) wins the selection decision. Activation at horizontal field location goes through the detection instability first and suppresses other field locations. **(B)** Same when the additional stimulus is applied on trial six. The stimulus at the novel location (bottom) wins the competition when activation becomes supra-threshold at vertical movement direction and other field locations are inhibited.

is chosen such that it may induce a supra-threshold peak in the intention field. Therefore it is larger than the familiar task input $s_T(\mathrm{H})$. The parameter values of the model were left unchanged.

Figure 10 shows the resulting time courses of activation in the two fields. When the second stimulus input at $x = V$ is applied on the third trial, activation at initial location, $x = H$, reaches positive values faster than at the novel location, despite the new input being larger than the familiar one. This is because peak formation at the familiar location is already facilitated by the stabilizing memory trace u_{mem} there (not shown). Because the intention field is selective, activation at $x = V$ then remains sub-threshold (see **Figure 10A**). When the second stimulus input at $x = V$ is applied on the tenth trial, activation at that new field location reaches the threshold faster than at the familiar location. This is because, at that familiar location, inhibition from the habituation field now predominates over the stabilizing memory trace (not shown). Once activation at $x = V$ goes through the detection instability other field locations are inhibited and activation at $x = H$ decreases (see **Figure 10B**).

In the A-not-B task, the selection decision to move to either the A or the B location is made on every trial. That selection is biased by the cue given to either location. Perseveration is measured as the preferred selection of the familiar movement even when the cue is given to the new movement. In the model, larger amplitude of the task input at $x = V$ may be interpreted as the cue given to that movement direction. The simulation shows that the model produces the same pattern, an early preference of the familiar choice, a late preference of the novel choice. The model thus unifies an account for habituation and perseveration for movement tasks.

4. DISCUSSION

We proposed a neural dynamic model that combines mechanisms previously postulated to explain perseverative motor behavior (Thelen et al., 2001; Dineva and Schöner, 2018) with mechanisms previously proposed to explain habituation to visual stimuli (Schöner and Thelen, 2006; Perone and Spencer, 2013b). This sets up an analogical mapping between the perceptual and motor domains. Perseveration in the motor domain corresponds to familiarity preference in the perceptual domain in that both are being caused by the build-up of activation in excitatory neural representations of movement parameters and of visual perceptions, respectively. The build-up of activation in an inhibitory layer of such a representation is the cause of habituation in the perceptual domain. Dishabituation and novelty preference result when a novel stimulus is presented after habituation has occurred to an earlier (familiar) stimulus. The analogical mapping predicts that similar effects of habituation and dishabituation should be observed in the motor domain. The mapping also predicts that novelty preference should be observed in the motor domain after habituation to a familiar movement.

We reported experimental evidence for the first part of this prediction. By applying the typical habituation paradigm to a motor task, we found a significant decrease of duration over which movements were performed and of the total movement path length during the habituation phase. When a new movement direction was enabled, we observed recovery of the movement time and path, an index of dishabituation. When the original movement direction was tested again, we observed signatures of Spencer-Thompson dishabituation. These results provide evidence for habituation in motor behavior that is specific to a particular movement, here probed by movement direction. We showed that the neural dynamic model accounts for all three signatures, habituation, dishabituation, and Spencer-Thompson dishabituation, through an approximate quantitative fit.

We provided theoretical evidence for the second prediction by simulating the model in a selection task. Activation was first induced for one value of the movement parameter by providing input at the corresponding location in the field. When this input was paired with an input at a competing location, the model selected the initial (familiar) location early during a sequence of habituation trials, but selected the second (novel) location late during the sequence of habituation trials. Mapped onto the A-not-B paradigm, the first pattern is consistent with perseveration after a small number of A trials (Wellman et al., 1986; Smith et al., 1999), the second pattern is consistent with reduced perseveration and enhanced switching to B after a larger number of A trials (Marcovitch et al., 2002; Marcovitch and Zelazo, 2006).

Together, the experimental and modeling results support a unified account in which motor behaviors and orientation responses are stabilized early during the experience of a motor behavior or a percept. With extended experience, the motor behavior or orientation response is destabilized, which promotes switching to alternate motor behaviors or re-orientation to alternate perceptual objects. This unified account is possible within the framework of Dynamic Field Theory because that framework postulates that all behaviorally significant neural states are attractors, whose stability prevents change. Transitions to new behavioral states are mediated by instabilities, the reduction of the attractors' stability. In DFT, enhanced stability comes from the accumulation of activation in excitatory populations that was modeled here by a memory trace, but that could also occur through the strengthening of synaptic connections from inputs to the excitatory populations. Conversely, reduced stability comes from the accumulation of activation in inhibitory population, likewise modeled by a memory trace here, but potentially taking the form of strengthening of synaptic connections from excitatory to inhibitory populations. The switch of activation state within the neural dynamic fields directly implements the decision to engage in a particular movement behavior or orientation response. Earlier work has established how such decisions can be directly coupled into a dynamics of fixation and gaze shift (Kopecz and Schöner, 1995; Perone and Spencer, 2013b) and into a dynamics of reaching movements (Schöner et al., 2019). In that respect, the account goes beyond earlier neural dynamic models that use overlapping ideas, in which levels of activation are mapped onto amounts of looking (Sirois and Mareschal, 2004) or probabilities of reaching to a location (Munakata, 1998).

The link between the build-up of excitatory/inhibitory activation and stability/instability offers a perspective on how processes of behavioral and perceptual exploration may be steered. This is a very broad topic that has been studied in many different settings. One notion that can be formalized mathematically (e.g., Kompella et al., 2017) is that "curiosity," assigning high value to behaviors or state that create much variance, may structure the exploration of a state space. At a high level, this notion may appear compatible with a Sokolovian idea of investing into behaviors, while they are novel, and turning away from them, when they become known. In our much lower level account, such behavior is ultimately always directed at objects (Ruff, 1986), framed as perceptual outcomes or as the targets of movement behavior. By modeling the Sokolovian idea of "turning away from" as a destabilization of the ongoing behavior or orientation, the neural dynamic account suggests that exploration emerges as other objects or behaviors compete with a now destabilized earlier choice.

This raises the question, at which level this competition takes place. We looked only at a very low level of movement representations, the direction of a lever movement. Similarly, models of visual habituation have invoked very simple feature spaces, over which neural representations are built (Sirois and Mareschal, 2002; Schöner and Thelen, 2006). In reality, behavioral choices may be made at the levels of action goals (Raab and Hartley, 2018), potentially linked to the possible outcomes of such action (Herbort and Butz, 2012). Outcomes are perceptual events that occur once an action has been performed. Our account is far from reaching such a level, but it may be worthwhile to think through the implications for the concrete paradigms we modeled.

At what level may the movement decisions have been made in the experiment we reported? The effect of visual habituation was minimized, so we do not think that it is the visual appearance of the lever or the perception of hand's movement that matter. We also found that habituation did not depend on the hand used. So it is not likely, that the level of motor actions for particular effectors matters. The perceptual outcome of movement was the music that played in response to the toddler's movement. Unfortunately, the experiment did not probe the role of that outcome dimension. Informally, we observed that toddlers were not interested in moving the lever without perceiving the music. In the model, we tested how the omission of the reward input that models how the outcome affects the habituation process: Trials without reward input do not contribute or contribute less to habituation. The model predicts that the rewarding outcome of an action influences the intention to move and through that, the process of habituation. Analogously, the toy-less version of the A-not-B paradigm (Smith et al., 1999) shows that perseveration does not necessarily depend on knowledge about the hidden toy. Movements were motivated by attracting the infants attention to identical visible objects (lids) at the two locations. In this view, any outcome that is interesting enough to elicit a movement may impact on perseveration and habituation. A concrete task for future work would be to lift the ideas of stabilization and destabilization discussed in this article to the levels of goal and outcome representation, which would open goal selection and outcome prediction to neural dynamic accounts. Empirical support for such a generalization may come from the paradigm of voluntary task switching (Arrington and Logan, 2004).

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because there is no agreement to share the data. Requests to access the datasets should be directed to gregor.schoener@rub.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Department of Psychology, Indiana University, Bloomington IN, USA. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the participants' legal guardian/next for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

JF and GS designed the experiment. JF performed the experiment and analyzed the data. SA and GS designed the model and wrote the manuscript. SA performed model simulations. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the Deutsche Forschungsgemein-schaft (DFG, German Research Foundation)—402791869 (SCHO 336/12-1) within the SPP The Active Self (SPP 2134). We acknowledge support by the Open Access Publication Funds of the Ruhr-Universität Bochum.

ACKNOWLEDGMENTS

The experiments were developed and performed in Esther Thelen's infant laboratory at Indiana University. JF and GS cherish the interaction they enjoyed over many years with the late Esther Thelen.

REFERENCES

Arrington, C. M., and Logan, G. D. (2004). The cost a voluntary task switch. *Psychol. Sci.* 15, 610–615. doi: 10.1111/j.0956-7976.2004.00728.x

Balkenius, C. (2000). Attention, habituation and conditioning: toward a computational model. *Cogn. Sci. Quart.* 1, 171–214.

Colombo, J., and Mitchell, D. (2009). Infant visual habituation. *Neurobiol. Learn. Memory* 92, 225–234. doi: 10.1016/j.nlm.2008.06.002. Infant

Colombo, J., and Mitchell, D. W. (1990). "Individual differences in early visual attention: Fixation time and information processing," in *Individual Differences in Infancy: Reliability, Stability, Prediction*. New York, NY: Psychology Press. 193–227.

Dineva, E., and Schöner, G. (2018). How infants' reaches reveal principles of sensorimotor decision making. *Connect. Sci.* 30, 53–80. doi: 10.1080/09540091.2017.1405382

Erlhagen, W. and Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychol. Rev.* 109, 545. doi: 10.1037/0033-295x.109.3.545

Goldberg, J., and Schöner, G. (2007). "Understanding the distribution of infant attention: a dynamical systems approach," in *Proceedings of the 29th Annual Cognitive Science Society*, eds D. S. McNamara, and J. G. Trafton (Austin, TX: Cognitive Science Society), 1043–1048.

Herbort, O., and Butz, M. V. (2012). Too good to be true? Ideomotor theory from a computational perspective. *Front. Psychol.* 3, 1–17. doi: 10.3389/fpsyg.2012.00494

Kompella, V. R., Stollenga, M., Luciw, M., and Schmidhuber, J. (2017). Continual curiosity-driven skill acquisition from high-dimensional video inputs for humanoid robots. *Artif. Intell.* 247, 313–335. doi: 10.1016/j.artint.2015.02.001

Kopecz, K., and Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biol. Cybern.* 73, 49–60. doi: 10.1007/BF00199055

Marcovitch, S., and Zelazo, P. D. (2006). The influence of number of a trials on 2-year-olds' behavior in two a-not-b-type search tasks: a test of the hierarchical competing systems model. *J. Cogn. Develop.* 7, 477–501. doi: 10.1207/s15327647jcd0704_3

Marcovitch, S., Zelazo, P. D., and Schmuckler, M. A. (2002). The effect of the number of a trials on performance on the a-not-b task. *Infancy* 3, 519–529. doi: 10.1207/S15327078IN0304_06

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a PDP model of the AB task. *Develop. Sci.* 1, 161–184.

Perone, S., and Ambrose, J. P. (2016). A process view of learning and development in an autonomous exploratory system. *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, NY: Oxford University Press. 271–296.

Perone, S., and Spencer, J. P. (2013a). Autonomous visual exploration creates developmental change in familiarity and novelty seeking behaviors. *Front. Psychol.* 4, 648. doi: 10.3389/fpsyg.2013.00648

Perone, S., and Spencer, J. P. (2013b). Autonomy in action: linking the act of looking to memory formation in infancy via dynamic neural fields. *Cogn. Sci.* 37, 1–60. doi: 10.1111/cogs.12010

Raab, H. A., and Hartley, C. A. (2018). "The development of goal-directed decision making," in *Goal-Directed Decision Making–Computations and Neural Circuits*, eds R. Morris, A. Bornstein, and A. Shenhav (New York, NY: Academic Press), 279–308.

Roder, B. J., Bushnell, E. W., and Sasseville, A. M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy* 1, 491–507. doi: 10.1207/S15327078IN0104_9

Ruff, H. A. (1986). Components of attention during infants ' manipulative exploration. *Child Develop.* 57, 105–114.

Schöner, G., Spencer, J. P., and DFT Research Group, T. (2016). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, NY: Oxford University Press.

Schöner, G., Tekülve, J., and Zibner, S. (2019). "Reaching for objects : a neural process account in a developmental perspective," in *Reach-to-Grasp Behavior: Brain, Behavior and Modelling Across the Life Span*, eds D. Corbetta, and M. Santello (Taylor & Francis), 281–318.

Schöner, G., and Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychol. Rev.* 113, 273–299. doi: 10.1037/0033-295X.113.2.273

Sirois, S., and Mareschal, D. (2002). Models of habituation in infancy. *Trends Cogn. Sci.* 6, 293–298. doi: 10.1016/s1364-6613(02)01926-5

Sirois, S., and Mareschal, D. (2004). An interacting systems model of infant habituation. *J. Cogn. Neurosci.* 16, 1352–1362. doi: 10.1162/0898929042304778

Smith, L. B., Thelen, E., Titzer, R., and McLin, D. (1999). Knowing in the context of acting: the task dynamics of the A-not-B error. *Psychol. Rev.* 106, 235–260.

Sokolov, E. (1963). *Perception and the Conditioned Reflex*. New York, NY: Pergamon Press.

Thelen, E., Schöner, G., Scheier, C., and Smith, L. (2001). The dynamics of embodiment: a field theory of infant perseverative reaching. *Brain Behav. Sci.* 24, 1–33. doi: 10.1017/s0140525x01003910

Thompson, R. F., and Spencer, W. A. (1966). Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychol. Rev.* 73, 16–43.

Wellman, H. M., Cross, D., and Bartsch, K. (1986). Infant search and object permanence: a meta-analysis of the A-not-B error. *Monographs Soc. Res. Child Develop. 214* 51, 1–67.

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

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

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: frontiersin.org/about/contact

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