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

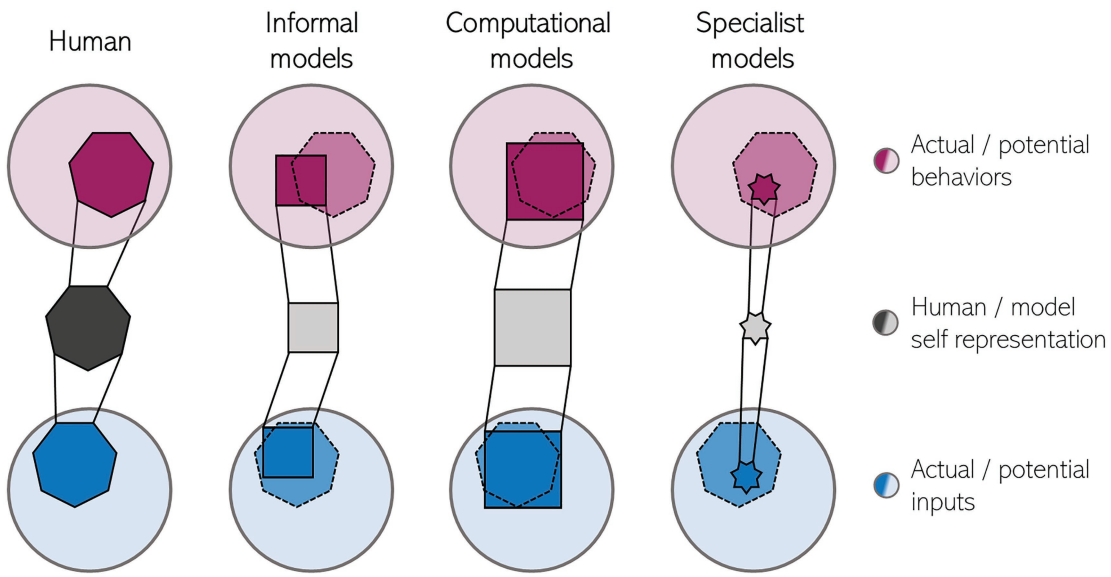

FIGURE 1 | Schematic of the representational power of models for the human minimal self. The human minimal self representation (dark grey, left) is based on a relevant subset of all self-related information (blue) while also underlying a subset of all possible self-related behaviors (purple, e.g., experiencing 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.