Computational models of the "active self" and its disturbances in schizophrenia

Tim Julian Möller, Yasmin Kim Georgie, Guido Schillaci, Martin Voss, Verena Vanessa Hafner, Laura Kaltwasser

▶ **To cite this version:**

Tim Julian Möller, Yasmin Kim Georgie, Guido Schillaci, Martin Voss, Verena Vanessa Hafner, et al.. Computational models of the "active self" and its disturbances in schizophrenia. Consciousness and Cognition, 2021, 93, pp.103155. 10.1016/j.concog.2021.103155 . hal-03344246

HAL Id: hal-03344246
https://hal.science/hal-03344246

Submitted on 14 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Consciousness and Cognition 93 (2021) 103155

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Review article

Computational models of the "active self" and its disturbances in schizophrenia

Tim Julian Möller [a,b,*], Yasmin Kim Georgie [c], Guido Schillaci [d], Martin Voss [a,e], Verena Vanessa Hafner [c], Laura Kaltwasser [a,b]

[a] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany
[b] Department of Psychiatry and Psychotherapy, Charité University Medicine, Berlin, Germany
[c] Department of Computer Science, Humboldt-Universität zu Berlin, Germany
[d] The BioRobotics Institute and Dept. of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy
[e] Department of Psychiatry and Psychotherapy, Charité University Medicine and St. Hedwig Hospital, Berlin, Germany

ARTICLE INFO

Keywords:
Schizophrenia
Self-disorders
Minimal self
Active self
Sense of agency
Sense of ownership
Computational psychiatry
Cognitive robotics
Developmental robotics
Predictive processing

ABSTRACT

The notion that self-disorders are at the root of the emergence of schizophrenia rather than a symptom of the disease, is getting more traction in the cognitive sciences. This is in line with philosophical approaches that consider an enactive self, constituted through action and interaction with the environment. We thereby analyze different definitions of the self and evaluate various computational theories lending to these ideas. Bayesian and predictive processing are promising approaches for computational modeling of the "active self". We evaluate their implementation and challenges in computational psychiatry and cognitive developmental robotics. We describe how and why embodied robotic systems provide a valuable tool in psychiatry to assess, validate, and simulate mechanisms of self-disorders. Specifically, mechanisms involving sensorimotor learning, prediction, and self-other distinction, can be assessed with artificial agents. This link can provide essential insights to the formation of the self and new avenues in the treatment of psychiatric disorders.

1. Introduction

Computational psychiatry is a recent approach that aims at explaining psychiatric disorders on a computational level. Specifically, different disorders are modelled within a complex cognitive system by looking at aberrant computations in the brain. One disorder that has been of interest is schizophrenia because of the effective translation of psychotic symptoms into a predictive coding framework (Sterzer et al., 2018; Heinz et al., 2019).

The World Health Organization defines schizophrenia as a "severe mental disorder characterized by profound disruptions in thinking, affecting language, perception, and the sense of self. It often includes psychotic experiences, such as hearing voices or delusions" (World Health Organization, 2001). Much research focus has been placed on its cognitive symptoms (Addington, Addington, & Maticka-Tyndale, 1991; Andreasen, Arndt, Alliger, Miller, & Flaum, 1995; Rector, Beck, & Stolar, 2005; Simpson, Kellendonk, &

* Corresponding author at: Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany.
E-mail addresses: tim.julian.moeller@hu-berlin.de (T.J. Möller), yasmin.kim.georgie@informatik.hu-berlin.de (Y.K. Georgie), guido.schillaci@santannapisa.it (G. Schillaci), martin.voss@charite.de (M. Voss), hafner@informatik.hu-berlin.de (V.V. Hafner), laura.kaltwasser@hu-berlin.de (L. Kaltwasser).

https://doi.org/10.1016/j.concog.2021.103155
Received 27 January 2021; Received in revised form 14 May 2021; Accepted 20 May 2021
Available online 12 June 2021
1053-8100/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Kandel, 2010). However, recently more researchers view schizophrenia as a disorder that is rooted in the disconnectedness from one's own body, which leads to experiences of a disturbed sense of self and intentionality, and in severe cases a complete loss thereof (Sass & Parnas, 2003; Sterzer, Mishara, Voss, & Heinz, 2016). Viewing schizophrenia as a self-disorder and analyzing it by means of computational approaches is thought to have explanatory value which could lead to more effective treatments of the disorder. One such novel approach is using virtual reality (VR) paradigms: When patients with schizophrenia see their own heartbeat exteroceptively in the form of visual input, it could help them to identify their own body processes that might ultimately lead to a stronger connectedness towards their own body, leading to a better treatment of self-disorders. In fact, a recent study using VR for a social attention task in patients with a psychotic disorder, revealed medium to strong effect sizes of symptom reduction (Adery et al., 2018).

Analyzing self-disorders poses a special challenge since a universal definition of "the self" is still lacking and there are different approaches in the examination of the self and how it emerges. Developing computational theories of the self can help us better understand the self and its minimal requirements, which may lead to new approaches for treating self-disorders. A definition of the self as a minimal form of self experience provides avenues for researching this phenomenal experience in an empirical setting (Gallagher, 2000). Additionally, a self model and the ability to distinguish between self and other in a robot would allow safer and more natural interactions with humans in complex and dynamic environments (Lanillos & Cheng, 2018; Hafner, Loviken, Pico Villalpando, & Schillaci, 2020). One such promising approach should place emphasis on the body and how the phenomenal experience of embodiment constitutes the emergence and the sense of self.

Research in humans as compared to animal models of schizophrenia (Marcotte, Pearson, & Srivastava, 2001; Jones, Watson, & Fone, 2011; Van Den Buuse, Garner, Gogos, & Kusljic, 2005) allows access to the subjective experience of schizophrenia in the form of self reports, but the underlying computations are in a "black box". Using embodied artificial agents to examine phenomenological concepts such as the self may provide access to the processes and flow of information in the cognitive system, and relate those to action and behavior. The interdisciplinary field of cognitive developmental robotics (CDR) might provide more insight to this "black box". CDR can be best understood as an intersection between robotics, developmental neuroscience, and developmental psychology which instantiates models of artificial agents on the basis of animal and human ontogenetic development (Lungarella, Metta, Pfeifer, & Sandini, 2003). CDR aims to understand the developmental processes of human cognitive functions and abilities such as self/other distinction in artificial agents through an "in silico" approach (Asada et al., 2009).

Finally, cognitive developmental robotics offers a testing ground for developmental theories and models about the self that are more accessible to study in artificial agents compared to biological beings, like the role of sensorimotor learning, internal body representations, and the re-enactment of sensorimotor experience in the development of skills, such as self-recognition (Asada et al., 2009; Weng et al., 2001).

In this review, we first provide an overview of current philosophical debates on the nature of the self in Section 1.1. In Section 1.2, we discuss schizophrenia as a disorder of self experience. Section 2 presents the leading computational approaches in modeling the self, as well as neuroscientific research into human self-models. In Section 3, the aim and guiding principles of computational psychiatry are described, its relevance in light of a selection of disorders, as well as the leading computational approaches in the field. Specifically, we highlight Bayesian-based approaches as promising avenues for modeling self-disorders, but also discuss some challenges. The field of cognitive developmental robotics is presented in Section 4 as providing a platform for implementing models of self-disorders in artificial agents. In Section 5, we analyze the common ground of both disciplines, and discuss why and how both fields can work together in order to verify and test models of the disordered self in schizophrenia. We argue for embodied artificial agents as perfect test-beds for scientific investigations of computational theories in psychiatry. Furthermore, new avenues for novel research topics such as hallucinations and dreamstates in robotic systems can be derived by merging both disciplines. We conclude in Section 6 and argue for advantages in using embodied robotic systems to simulate and gain insight into computational mechanisms underlying neurocognitive processes and self-disorders.

1.1. Philosophical approaches in defining the self

The concept of the self has been examined within numerous disciplines and perspectives. At the heart of the philosophical discussion about the self are the subjective experience of selfhood and consciousness. The subjective experience and the self are usually described as having conscious and unconscious, reflective, and pre-reflective aspects, or as having levels of organization. For example, a phenomenological view by Parnas and Handest (Parnas & Handest, 2003) describes selfhood on three hierarchically-organized but entwined levels that go from a first personal givenness of experience, "an implicit, pre-reflective egocentricity determining the very manifestation of experience…experience and self are not separate entities. Rather, the first personal perspective is a way in which the experience articulates itself." (Parnas & Handest, 2003, p. 122), to more elaborated and sophisticated levels of describing selfhood, such as self-awareness, or a social self (Parnas & Handest, 2003).

The most basic experience of being a self, minimal phenomenal selfhood (Blanke & Metzinger, 2009), has been described as a "non-conscious experience of being bounded within a sentient body" (Riva, 2018, p. 245). Other low-level concepts of self include the unconscious proto-self (Damasio, 1999) consisting of sensorimotor representations of the body in interconnected brain regions, and the immunological self, consisting of "what the immune system identifies as belonging to the body" (Damasio, 2003, p. 227).

Gallagher (2000) offers to define a minimal form of the self as the "consciousness of oneself as an immediate subject of experience, unextended in time" (Gallagher, 2000, p. 15). Phenomenologically, this *minimal self* consists of a sense of agency—the self as the one causing or generating an action, and the sense of ownership—the sense of the self as the one subjected to an experience (Gallagher, 2000; Van Den Bos & Jeannerod, 2002). The adjective *pre-reflective* stands for an experience that comes before "the moment one attentively inspects or reflectively introspects" (Gallagher & Zahavi, 2016).

These conceptualizations of the self on the unconscious, pre-reflective level, highlight the importance of the body and embodiment in constituting the self, its emergence, and development. The concept of phenomenal transparency is also involved in the discussion of the self and its embodiment, aiming at an explanation on why we do not have a conscious experience of our body and senses. For example, Metzinger (2014) proposed that we cannot experience our self-models interoceptively as such, because the self model is phenomenologically transparent or in other words, consciously non-accessible. According to Metzinger's theory of subjectivity, our experience of being a self is an illusion caused by its phenomenologically inaccessible vehicle properties.

Thompson and Stapleton (2009) suggest that we experience the world through the "transparent body", as a pre-reflective bodily experience, and that "environmental resources that are incorporated gain this transparency. They are no longer experienced as objects; rather the world is experienced through them" (Thompson & Stapleton, 2009, p. 29). The enactive approach (Paolo & Thompson, 2014) relates this transparency to sense-making: "we propose the following 'transparency constraint': For anything external to the body's boundary to count as a part of the cognitive system it must function transparently in the body's sense-making interactions with the environment." (Thompson & Stapleton, 2009, p. 29). Therefore, in this approach, "tools and aids that conform to transparency are incorporated into the neurophysiological body schema" (Thompson & Stapleton, 2009, p. 29).

According to Northoff (2013), the self model is based on summarizing, integrating, and coordinating the information flow between the body and brain. Northoff distinguishes four concepts of the self: The mental self, the empirical self, the phenomenal self, and the minimal self. Other notions are the ecological self, which describes the self being perceived with respect to the physical environment as "I am the person here in this place, engaged in this particular activity" (Neisser, 1988, p. 245), and the embodied self as "I am the owner of a body and the author of actions" (Jeannerod, 2007, p. 126). The mental self is based on our thoughts and a "specific mental substance", while the empirical self is rather based on representing and reflecting about own biological processes, than a mental substance. The phenomenal self represents conscious experience that comes with (pre-reflective) self-awareness. Lastly, the minimal self is defined as having its basis in our body and its related physiological processes (Northoff, 2013). As argued by Christoff, Cosmelli, Legrand, and Thompson (2011), these three building blocks should be complemented by the fundamental, but often neglected, experience of oneself as an agent. This crucial aspect of self-experience ultimately gives rise to the feeling of being a "cognitive-affective" agent. The Phenomenal Self (PS) is widely considered to depend on body representations and as a locus of subjectivity that is necessary for all conscious experience. Of crucial importance is the body as a source of sensory and motor information (e.g. vision, touch, proprioception) (Gallagher, 2000). A model of selfhood must thereby go beyond the mere subjective experience and integrate interoceptive and exteroceptive signals over multiple levels of self-representation, i.e. physiological homeostasis, physical bodily integrity, morphology, and position (Seth, 2013).

The narrative self is a more elaborated aspect of the self. It is selfhood in its explicit, conceptual form, emerging through interactions with others, and specifically through identifications of selfhood underlain by language and culture (Gallagher, 2000).

In approaching to answer how the self manifests itself, it is not sufficient however, to only look at ourselves alone; the fundamental experiences of our bodily selves are shaped from the very beginning by the physical, as well as the mental interactions like mirror mechanisms, with other bodies (Gallese & Sinigaglia, 2010). The body is not only existent in the current world, but the worlds' affordances are a source of motor potentialities that can alter the world we live in. In order to explain the self, not only the attunement of inanimate things, but foremost an attunement with other self-modeling bodies has to be constituted (Gallese & Sinigaglia, 2010). The sense of self can thereby be seen as a transient process (Prescott & Camilleri, 2019), that evolves into an "object of consciousness, when prediction error for the actional outcome momentarily increases" (Tani & White, 2020, p. 3).

With regards to how the self emerges and constructs itself, there has been an evergrowing focus on the enactive approach that accentuates the role of active movement for constituting a mind. It follows that consciousness and the process of generating a "self" is closely entangled with agency, intention, and behavior (Hohwy & Seth, 2020; Hommel, 2013; Ma & Hommel, 2015b). The enactive approach postulates that cognition emerges in the interactional domain between the agent and the environment as a process of sense-making through action. Therefore, the world is not merely passively "represented" in the agent's mind but rather constituted by the embodied relational process of interacting with the world. In other words, the agent does not passively receive information that is later represented, but rather the agent enacts a world through sense-making (Thompson & Stapleton, 2009; Paolo & Thompson, 2014).

Following the enactive approach, we refer to the notion of the "active self" of an embodied agent, as a construct which intersects prediction, action, and perception in interactions with the environment. It is within this intersection, that the sensorimotor activities of the agent are thought to facilitate the emergence of a minimal self, such that "the phenomenal, minimal self is empirically derived from sensorimotor experience and not a theoretical and empirical given" (Verschoor & Hommel, 2017, p. 140). The active self participates in an iterative, enactive process: active interaction with the environment affects perception (Thompson & Stapleton, 2009; Paolo & Thompson, 2014), giving rise and modifying phenomenal experiences and internal states, thereby actively participating in its construction (Nguyen et al., 2020). The concept of the active self is therefore relevant to the view of schizophrenia as a self-disorder, which is thought to be rooted in disconnectedness from one's body (Sass & Parnas, 2003), and often entails an altered sense of agency (Jeannerod, 2009; Voss, Chambon, Wenke, Kühn, & Haggard, 2017; Voss et al., 2010).

1.2. Schizophrenia as a self-disorder

The World Health Organization (WHO) characterizes schizophrenia as involving disruptions in cognition, perception, as well as to the sense of self, and often entails psychotic experiences such as hallucinations or delusions (World Health Organization, 2001). The notion that self-disorders are at the root for the emergence rather than a symptom of schizophrenia and its related spectrum disorders is gaining more traction (Sass & Parnas, 2003; Parnas & Handest, 2003; Møller & Husby, 2000). These self-disorders can manifest themselves in a reduced sense of self, even to a point of retreating to complete automatism, in which patients disconnect completely

from observing and thinking (hyperautomaticity) (De Haan & Fuchs, 2010). In some patients this can lead to solipsistic delusions, in which an individual ontological reality exists (Parnas & Sass, 2001), or the existence of other sentient beings and even the external world outside one's conscious experience is denied completely (Bradley, 2016). For clinical case studies, see Parnas and Sass (2001). On the other extreme, self-disorders can also manifest in exaggerated sensation monitoring and self-consciousness (hyper-reflexivity) (Sass & Parnas, 2003). However, recent evidence suggests the experiences of hearing voices might be more common in the non-help-seeking population. Especially in the clairaudient psychics community, evidence suggest they were less distressed by their auditory hallucinations and the reception from their peers about their experience was more likely to be positive, leading to a lesser disruption of their social relationships. Interestingly, psychics showed to have more agency over their auditory hallucinations. Specifically, they were able to control the onset and offset of their voice-hearing experience (Powers, Kelley, & Corlett, 2017). Overall, this implies the continuum from health to disease might be larger than suggested by the WHO. While most psychiatrists treat schizophrenia as a mainly cognitive disorder, a growing body of literature views schizophrenia as a disorder rooted in a disconnectedness from one's body and a lack of intercorporal attunement that ultimately leads to the loss of self and intentionality. Recognizing the fundamental role of embodiment is therefore crucial in understanding schizophrenia (De Haan & Fuchs, 2010; Fuchs, 2005). In fact, a phenomenological exploration in patients suffering from schizophrenia reveals multiple layers of disconnectedness from the world and a breakdown of their perception of the world into separate chunks rather than a synthesis thereof (De Haan & Fuchs, 2010; Stanghellini, 2004). This points to the importance of analyzing subjective experiences when investigating self-disorders.

In order to regain the connection to the world, Sass and Parnas (2003) argue, hyperreflectivity and hyperautomatic behavior are used as a coping strategy of a disembodied mind rather than it being a mere symptom (Sass & Parnas, 2003; Sass, 2004; Sass & Parnas, 2007). However, there is still an ongoing debate on the extent to which these mechanisms remain conscious or unconscious; while it has been argued that schizophrenia can be described as a deficiency of automatic processes that lead to aberrant contents of consciousness, others argue for a clear distinction between them. Several authors hypothesized that schizophrenia is characterized by a failure of automatic processing leading to abnormal contents of consciousness (Gray, Feldon, Rawlins, Hemsley, & Smith, 1991; Maher, 1983; Frith, 1979). For example, it has been proposed that features that are usually unconscious (e.g. physiological processes, automatisms) glide into conscious awareness in patients with schizophrenia, due to aberrations in information processing (Gray et al., 1991). On the other hand, alternative approaches focus on unconscious dynamics that are distinct from conscious processes, and explain self-disturbances such as a decreased feeling of having a minimal self as being linked to general impairments in perception (Giersch & Mishara, 2017).

The rather new approach of computational psychiatry aims to characterize these mental disorders through multi-leveled aberrant neuronal computations (Montague, Dolan, Friston, & Dayan, 2012). By investigating aberrant computations in the brain, we could describe psychiatric disorders, disruptions of the self, and the senses of ownership and agency through these computational alterations.

2. Computational models of perception, action, and the self

Many ideas of modern computational psychiatry and cognitive robotics are based on the ideas of Bayes theorem (Bayes, 1763) and Helmholtz's unconscious inference (Helmholtz, 1867). According to Helmholtz, the motor organ serves as a tool for exploration in order to deduce invariances by trial and error, thus deriving real world knowledge (Westheimer, 2008). In the following sections, the adaptation of different approaches in computational psychiatry will be described.

2.1. The Bayesian brain hypothesis

A growing body of literature (Barber, Clark, & Anderson, 2003; Khrennikov, 2004; Rao, Olshausen, & Lewicki, 2002) points to the notion that the brain represents sensory information in the form of probability distributions rather than in a deterministic "as it is" manner (Knill & Pouget, 2004).

In short, a "Bayesian brain" continuously improves its fit by updating its model to minimize the error (Friston, 2012; Friston, 2003; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). In Bayesian terms, perception can be regarded as the brain's inference about the origin of sensory information that is calculated by comparing sensory information with predictions based on their priors (Jardri & Deneve, 2013). Specifically, this Bayesian brain model consists of three distributions: the prior, the likelihood, and the posterior. Each distribution can be thought of as a probabilistic Gaussian curve. The prior refers to the model's prediction, in other words the expectations of an agent. The likelihood refers to the sensory input. The posterior can be seen as the perception that represents a compromise between the prior and the sensory evidence. Each of the three distributions can be altered in precision, i.e. through clinical conditions like schizophrenia, or certain drugs. Conceptually, precision can be thought of as a steeper or flatter Gaussian curve (alpha), that influences the probability of e.g. the sensory evidence that would ultimately influence perception. A mismatch between the prior belief and the likelihood results in a prediction error that can be used to update the model's priors. The higher the inverse variance of the likelihood, the higher the error signal will be, and thus more influential in updating the brain's model (Adams, Brown, & Friston, 2014). However, if the prior also has a high precision, the prior is computationally more resistant to being updated. The brain is therefore required to balance the precision of the priors and likelihood to minimize prediction errors (Adams et al., 2014; Fletcher & Frith, 2009). Under these assumptions, schizophrenia can be understood as the result of an imbalance of prior and likelihood precision (Humpston & Broome, 2020; Horga & Abi-Dargham, 2019; Stephan & Mathys, 2014). This model will be discussed in more detail in Section 3.2.

2.2. Predictive processing models

Based on Bayesian statistics, the "predictive coding framework" and its hierarchical application in a cognitive context, the predictive processing approach, were developed. These postulate that the minimization of free energy is approximately equivalent to the maximization of model evidence, which corresponds to the maximization of the mutual information between sensory input and internal representations (Friston, 2012; Friston, 2005). Simply put, the brain engages in testing the accuracy of its internal model by formulating hypotheses for the sensory origin of its sensation, and resolving discrepancies by either revising its model, or by changing the bottom-up information to match the predictions (Limanowski & Blankenburg, 2013).

The predictive coding framework is nowadays widely used for research in many fields such as robotics, data science, neuroscience, and computational psychiatry. The strength of this framework is that it can be employed as a "common language" for modeling approaches, which facilitates interdisciplinary research. Also, with the help of the predictive coding framework, closer links between the phenomenology of the self and conscious experiences, and their mechanistic properties of neural substances can be analyzed (Hohwy & Seth, 2020). Specifically, the assumptions of predictive coding are well represented in the neuroanatomical structure of the cortex, indicating that the encoding of neuronal populations is indeed probabilistic (Clark, 2013; Friston, 2012; Bastos et al., 2012). As reviewed in detail by Hohwy and Seth (2020), the most promising approaches in the science of consciousness and the self indeed have uncertainty reduction and top-down signalling as a major foundation (e.g. integrated information theory, global neuronal workspace theory, recurrent processing theory; for a review, see Hohwy & Seth, 2020). Specifically, it seems that conscious systems have a tendency to settle in one unified, highly informative representational state (and maintain this homeostasis), by reducing uncertainty through learning and information integration (Tononi, Boly, Massimini, & Koch, 2016).

Furthermore, it seems unlikely that a self can emerge without top-down signalling, complementing bottom-up signalling, shown e.g. by anesthesia studies that indicate a disruption of top-down signal loops (Boly et al., 2012). Areas in which predictive coding surprise minimization is thought to take place are perception, action, attention, recognition, understanding, and exploration. Specifically, while the internal model always undergoes updates (perception), this process can be guided through active inference. Concretely, by engaging motor actions, specific sensory evidence can be acquired (action). Moreover, policies for behavior can be selected with the goal to reduce the expected prediction error. Goals can thereby be seen as priors within the predictive coding framework, guiding the action command of an agent. Sensory evidence can be further modulated by attention guided alterations of precision. Specifically, prediction error minimization is affected by the precision of sensory evidence, that can be increased by focusing on a specific target (attention). Lastly, following the Bayesian notion, model complexity is penalized which ultimately leads to simpler models in which prediction errors can be minimized as in the case of recognition, understanding, exploration.

It is hypothesized that by these statistical inferences, the brain generates an embodied self-model that on the long run, behaves as a representational system. Since the error dynamics of the world are ever changing, this system has to be capricious to maintain homeostasis, meaning it has to be hierarchically constructed, and capable of forming meta-expectations (Hohwy & Seth, 2020).

2.3. The comparator model

In order to have a subjective experience of oneself in the world, it is crucial that we can implicitly draw a boundary between ourselves and others. While the source of one's movement is important for self-identification, self-other distinction in terms of action intention attribution requires goal orientation in order for an agent to understand and attribute actions of others (Jeannerod, 2007). Tsakiris (2010) gives a neurocognitive model for the development of ownership (Tsakiris, 2010). According to this model, ownership develops through the interaction of multisensory input and multiple self-related internal models of the agent. First, a model of one's body enables the distinction between that which belongs to one's body, and that which does not. Moreover, the representation of the location of own body parts in space modulates the sensory information which might lead to a recalibration of the visual and tactile positions. Ultimately, the resulting coordinate system of tactile sensations leads to a subjective experience of body ownership (Tsakiris, 2010).

According to the comparator model, whenever the brain initiates a new movement, an internal prediction model for this movement is generated which is subsequently compared to the actual movement achieved. If there is a match between the actual movement and its prediction, action authorship is more strongly attributed to one's own body movement and thus sense of agency is perceived (David, Newen, & Vogeley, 2008). While the comparator model is an approach that is often used as an objective indicator to describe whether an agent subjectively experiences agency based on a mismatch between predicted and actual action (i.e. a low prediction error is interpreted as indicating high sense of agency), this approach has been criticized as being too simplistic (Lanillos, Pagès, & Cheng, 2020b; Zaadnoordijk, Besold, & Hunnius, 2019; Synofzik, Vosgerau, & Newen, 2008). Some authors stressed the more complex interplay between sensory input and efference copy signals (Synofzik, Vosgerau, & Voss, 2013). More specifically, the comparator model has been criticized as lacking important top-down components, i.e. sensorimotor contingency detection and causal inference, that are important for a more realistic model of self-detection (Lanillos et al., 2020b). Based on a model by Wegner (2017), Lanillos et al. (2020b) proposed to extend the current comparator model to the "double comparator" model that also takes processes like spatio-temporal contingency into account, to perform a distinction between the self and other.

Motor commands activate neuronal discharges—the efference copy or corollary discharge—that affects activity in both sensory and motor pathways, allowing an organism (fish, insects, mammals, etc.) to monitor and if necessary alter motor activity before muscle contraction actually takes place. Furthermore, it gives an agent the information whether a movement is self-generated or caused by an external source, leading to a successful mechanism of self-other distinction. According to this theory, sensory illusions experienced by patients with schizophrenia (and their alterations in self-other distinction) could result from a disordered internal feedback loop to the

high level sensory representations. Since most drugs given to treat symptoms of schizophrenia are capable of producing extrapyramidal syndromes (motor disorders such as dystonia, akathisia, parkinsonism, and in some cases tardive dyskinesia), they might alter internal striatal pathway communication (Feinberg, 1978). Specifically, efference copies are sent via pyramidal tract neurons to the dorsal striatum, synapsing with inhibitory GABAergic neurons (Fee, 2014; Shipp, 2017). The hypothesis is that excessive dopaminergic signalling leads to a stronger inhibition of the striatal transmission of the efference signal, thus impeding the internal monitoring of one's movement (efference copy), and producing extrapyramidal syndromes (McCutcheon, Abi-Dargham, & Howes, 2019).

This efference copy model has also been proposed as an explanation for why we cannot tickle ourselves. By attenuating self-produced tactile sensations through accurate sensory predictions of a forward model, we interpret this sensation as being caused by ourselves and thus as non-harmful, while if this sensory evidence was not predicted (and therefore surprising), we might be ticklish. This predictive mechanism is thought to be aberrant in patients with hallucinations or passivity experiences (e.g. schizophrenia) who have been found to be able to tickle themselves (Blakemore, Wolpert, & Frith, 2000; Pynn & DeSouza, 2013). Similar to the corollary discharge, the (double) comparator model has been proposed to explain sense of agency. Whenever the brain initiates a new movement, an internal prediction model for this movement is also generated that is then compared to the actual movement. If there is a match between the movement and its prediction, action authorship is more strongly attributed to one's own body movement and thus sense of agency is perceived (David et al., 2008). A graphical depiction of the classic comparator model is shown in Fig. 1.

2.4. Neuroscientific research on human self-models

There is ongoing discussion about the influence of top-down vs. bottom-up processes on sense of agency. While bottom-up approaches describe sub-systems that lead to the emergent property, top-down approaches gain insight by breaking down a system into its sub-systems.

Top-down influences such as psychotherapy, and bottom-up approaches like pharmacological interventions can substantially modify subjective experience (Fuchs, 2009). Famous examples of drugs that can lead to profound self-transformations are classic psychedelics like lysergic acid diethylamide (LSD), dimethyltryptamine (DMT), psilocybin, and mescaline (Baumeister & Exline, 2002; Preller et al., 2019; Timmermann et al., 2018; Mason et al., 2020; Hermle et al., 1992), but also clinically used substances such as ketamine (Vlisides et al., 2018). Nonetheless, it seems most likely that we perceive the world rather as top-down predictions of the world. Predictions that are then fine-tuned through bottom-up sensory experience (Hohwy & Seth, 2020). Ultimately, it is the interaction of bottom-up and top-down processes that shapes the beliefs and percepts which cannot be attributed to a single factor alone. Specifically, there is evidence that Glutamate may mediate top-down signals at NMDA receptors, and bottom-up signals at α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors while its integration may be mediated by dopamine (Wilson, Humpston, and Nathan, 2021; Sterzer et al., 2018).

On the other hand, much research has been conducted to investigate what underlies the self-model in terms of neuronal activity and brain function. Evidence for a strong top-down influence of certain brain areas that mediate the emergence and constitution of the self comes from ample optogenetic, transcranial magnetic stimulation (TMS), electrophysiological, and brain imaging studies. Some examples will be described below.

By using electrical pulses or light impulses, specific areas in the neuronal chain can be (de-) activated, thus giving the possibility to empirically investigate the self. Of interest is for example the Cortical Midline Structure (CMS) which is involved in mechanisms of self-reflection. Specifically, if the CMS is damaged, patients show impairments in evaluating problems they encounter and an

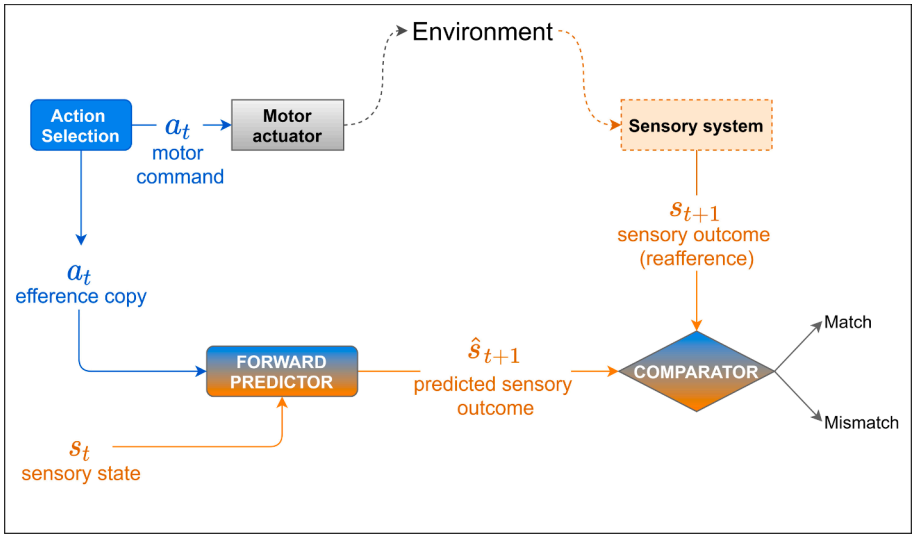

Fig. 1. The comparator model. This model explains the sense of agency as resulting from a match between an agent's internal movement prediction and its actual movement outcome.

overestimation of their performance and their capacities (van der Meer, Costafreda, Aleman, & David, 2010).

Recent studies using Electroencephalography (EEG) tried to disentangle neuronal bottom-up and top-down loops that are required for perception. Specifically, frequency-tagging could allow to unravel loops for prior expectation and attention. Moreover, they could show that expectation and attention increases perceptual top-down and bottom-up integration (Gordon, Tsuchiya, Koenig-Robert, & Hohwy, 2019).

Many forward and backward loops between neuronal populations in the human visual cortex share the same frequency. Microcircuits along feedforward and feedback projections in the human visual cortex show synchrony in their α, β, and γ-frequency bands, which is in line with circular inference models of brain dynamics (Leptourgos, Denève, & Jardri, 2017; Michalareas et al., 2016). Furthermore, semi-synchronous γ-frequency (40–70 Hertz) between neuronal populations pose a potential candidate as a neural correlate of consciousness (Michalareas et al., 2016). Also magnetoencephalography (MEG) experiments show evidence for the influence of probabilistic top-down priors on perception (Aru, Rutiku, Wibral, Singer, & Melloni, 2016).

The sense of self is a system-wide phenomenon thought to involve many brain regions with varying degrees of interconnectedness (Northoff et al., 2006; Tsakiris, Hesse, Boy, Haggard, & Fink, 2007; Knyazev, 2013; Tsakiris, 2017). Therefore, by inducing disturbances of the self through stimulating the implicated brain areas, one can pinpoint locations in the brain that are arguably involved in the sense of self. Evidence from TMS studies shows that different body parts are represented along the motor cortex (homunculus). By stimulating these areas, certain movements can be forced, e.g. the twitching of a finger or involuntary arm movements (Barker, Jalinous, & Freeston, 1985; Ziemann, Wittenberg, & Cohen, 2002). In another TMS study, Blanke et al. (2005) showed that stimulating the temporoparietal junction (TPJ) impairs mental transformation of the bodily self. Also, evoked potential mapping revealed that the TPJ is active 330–400 ms after stimulus onset when participants were asked to imagine themselves in the position and visual perspective that is usually reported by people reporting Out-of-body experiences (OBE). In a case-study of an epileptic patient with OBEs originating from the TPJ, partial activation of the seizure area during mental transformation of her body and visual perspective was seen. However, activation of other cortical sites was not found (Blanke et al., 2005). Therefore, it is suggested that the TPJ is a central brain area for conscious non-corporal self-experience, mediating spatial unity of self and body, while damage to this region could lead to pathological (temporal) loss of self such as in an OBE (Blanke et al., 2005; Blanke, Landis, Spinelli, & Seeck, 2004; Blanke, Ortigue, Landis, & Seeck, 2002). In another case-study by Blanke et al. (2002), the authors succeeded to repeatedly invoke corporal (as opposed to non-corporal visual hallucinations of the TPJ) OBEs in a patient undergoing epilepsy treatment by electrically stimulating the patient's right angular gyrus. Results suggest that the gyrus angularis could be a crucial node in a large neural circuit, involved in mediating complex own-body perceptions. Also, the experience of dissociation from ones body arises through a failure of the integration of complex somatosensory and vestibular information. These findings are in line with other studies that induced the feeling of a "proximal sentient being" and sensed presence within the laboratory. By applying pulsed magnetic fields over the temporoparietal region while wearing opaque goggles in a quiet room, a sense of presence and an experience of "another consciousness" could be induced in two thirds of the participants. This indicates that an altered sense of self, as well as ephemeral phenomena, like visitations by so called "spirits" or "gods" have a top-down guided Neuronal Brain Correlate of Consciousness (NCC) (Persinger & Healey, 2002).

Recent fMRI research in a psilocybin study indicates that glutamate might play an important role for ego-dissolution that is common in psilocybin experiences. Specifically, higher levels of medial prefrontal cortical glutamate were associated with negatively experienced ego-dissolution, while lower levels of hippocampal glutamate were associated with positively experienced ego dissolution (Mason et al., 2020). Additionally, gamma-aminobutyric acid (GABA) concentration deficits are found in the occipital cortex of patients with schizophrenia that are correlated with impaired visual inhibition (Yoon et al., 2010). The effect of GABA is specifically the inhibition synapses and its deficit in patients with schizophrenia is believed to cause the cognitive impairments in patients with schizophrenia (Cho, Konecky, & Carter, 2006). Multiple studies indicate that the posterior insular cortex is one central node for interoception and interactions with motor, somatosensory, and limbic systems (Ebisch & Gallese, 2015; Augustine, 1996; Craig, 2002; Craig & Craig, 2009) and probably contributes to self-awareness (Tsakiris et al., 2007). Specifically the neuronal activation of the posterior insular cortex is positively associated with the experience of the "rubber hand illusion" (Tsakiris et al., 2007). In this famous illusion, a rubber hand is placed in front of the participant, while their real hand is hidden from view. If the rubber hand and the participant's own hand are stroked synchronously with a brush, a multimodal conflict is induced which might lead to the experience that the rubber hand feel like one's own hand. Studies investigating an impaired sense of agency in patients with schizophrenia revealed an aberrant activation of the posterior insular cortex (Farrer et al., 2004). In a computational sense, the rubber hand illusion is thought to result from the integration of visual, tactile, and proprioceptive information and can be explained by the inference of a common cause thereof within a Bayesian causal inference with optimal multisensory integration (Samad, Chung, & Shams, 2015). This points to the notion that computational aberrations in patients with schizophrenia may lead to an altered experience of disembodiment and depersonalization. In fact, active inference as well as predictive processing have already been put forward as a promising account of symptoms of depersonalization and disembodiment and as a model for enacted existence in general (Deane, Miller, & Wilkinson, 2020; Seth, Suzuki, & Critchley, 2012; Gerrans, 2019; Hesp et al., 2021). Recently, emotions and "valenced bodily feelings" have also been proposed to represent a feedback source of information about the predictive success of an agent that all fundamentally shape the disturbances of the "minimal self" (Gerrans, 2019; Deane et al., 2020; Hesp et al., 2021).

Overall, while there can not be a single brain area dedicated to such a complex phenomenon as the self, it appears that cortical midline structures, temporoparietal junction, angular gyrus, and the insula play a major role in the emergence of phenomenal agentive and bodily self-experience such as spatial unity, self location, and egocentric visuo-spatial perspective (Blanke et al., 2005; Blanke et al., 2004; Richer, Martinez, Robert, Bouvier, & Saint-Hilaire, 1993; Jeannerod & Johnson-Frey, 2003; Ruby & Decety, 2001; Decety & Sommerville, 2003; Vogeley & Fink, 2003).

3. Computational psychiatry

3.1. Guiding principles of computational psychiatry

The rather new field of computational psychiatry formalizes brain functions mathematically or computationally to characterize mechanisms of psychopathology (Friston, Stephan, Montague, & Dolan, 2014). Computational psychiatry aims to deliver explanations especially of aberrant mental conditions through methods of calculations (Montague et al., 2012). In other words, the approach is to examine psychiatric conditions by looking at the disruptions in information flow and to gather knowledge about the principle governing brain function with the help of computational and mathematical models.

The incentive is that psychiatric conditions such as schizophrenia can be better understood by artificially altering computational models. Especially with newer theories such as the free energy principle and the Bayesian brain hypothesis, these mechanisms can be examined more directly in artificial agents, where the calculations are more easily accessible, compared to biological agents.

The sensible link between computational psychiatry and cognitive developmental robotics arises from the fact that both fields are concerned with the information flow in a complex cognitive system. Using cognitive developmental robotics provides the advantage of using a physical body. First, embodiment is important not just because it is closer to reality, but especially for self-disorders, since it is believed that self-disorders arise from a fundamental disconnectedness from one's body, rather than being a mere symptom.

Second, the developmental aspect is crucial in that a self is arguably an emergent property coming from developmental processes (Wolputte, 2004; Piaget, 1954; Erikson, 1950; Rochat, 2003). Therefore, one needs to have a developmental approach when modeling self-disorders in a robot. While schizophrenia is mainly diagnosed as a mental disorder characterized by its symptoms of altered perception, thoughts, mood, and behavior (National Collaborating Centre for Mental Health, 2014), its behavioral prodromes, felt disconnectedness towards one's body, and alterations in brain physiology are seen much earlier. In fact, the neurodevelopmental model of schizophrenia posits that the symptoms of schizophrenia are the end state of an aberrant neurodevelopmental process rather than a degenerative process (Rapoport, Giedd, & Gogtay, 2012; Murray & Lewis, 1987; Weinberger, Berman, & Zec, 1986; Insel, 2010; Owen, O'Donovan, Thapar, & Craddock, 2011). Even prenatally, the placental pathology is an indicator for the risk of developing schizophrenia (Rapoport et al., 2012). Therefore, these aberrant developmental processes can be simulated more thoroughly by using cognitive developmental robotics as opposed to non-embodied simulations.

An example where computational psychiatry is especially interesting is the case of apperceptive agnosia. Patients with apperceptive agnosia have an intact visual field and functional and consciously perceived low-level visual perceptions, but fail to recognize objects they are looking at, distinguish between different shapes, or copy the same shape. This disorder is often a result of selective lesions in the occipital and temporal cortex caused by a lack of oxygen or carbon monoxide poisoning (Heider, 2000). Even though patients with apperceptive agnosia have an integrated and coherent world-model, certain gestalt cues for organizing their visual perception are lacking (Metzinger, 2014). There is more evidence from disorders like autopagnosia in which patients cannot name, identify, or even localize their own body parts, which is also caused by cortical brain damage. Another interesting clinical picture where multimodal integration fails is disjunctive agnosia. Patients with this disorder cannot merge their visual with their auditory sensory input (Metzinger, 2014).

By taking a computational approach, these syndromes emerge through a disconnectedness of certain neural pathways and corresponding alterations of prior beliefs of underlying likelihood distributions (Parr, Rees, & Friston, 2018). This approach can provide explanations for various clinical conditions, e.g. the phantom limb syndrome (Frith, Blakemore, & Wolpert, 2000; De Ridder, Vanneste, & Freeman, 2014). In the phantom limb syndrome, patients who underwent an amputation of a limb still experience phantom sensory percepts, in some cases even pain. One of the most commonly cited explanations for this syndrome is cortical reorganization

Table 1
Marr's three level computational approach of information processing tasks.

	Computational theory	Representation and algorithm	Hardware implementation
Framework	What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?
Schizophrenia	Phenomenological anomalies of selfhood e.g. free energy principle framework	Aberrant predictive coding and active inference	Neurobiological anomalies (e.g. dopamine, glutamate, GABA) Neurophysiological anomalies (e.g. gray matter loss, enlarged lateral brain ventricles)
Robotics	Bayesian inference of the latent causes of sensory evidence	Predictive coding	artificial neural networks and using the robots physical body, e.g. arms, legs, sensors.

Marr's three levels for machine information processing: The computational level refers to what the device is doing and why. The algorithmic level to how input, output, and transformations are represented. The implementational level refers to how to processes are implemented physically. Adapted from Marr (1982).

(Baron, Binder, & Wasner, 2010; Ramachandran, Brang, & McGeoch, 2010; Flor, Nikolajsen, & Jensen, 2006). One non-pharmacological treatment is the mirror therapy (Ramachandran & Rogers-Ramachandran, 1996). In this treatment, a mirror is placed parasagittally between the intact and the missing limb and the patient is instructed to move the missing, and the intact limb in the same manner. Thereby, the patient sees the phantom limb movements as reflection in the mirror. This movement might resolve the visual-proprioceptive dissociation in the brain and might ultimately lead to a symptom reduction (Ramachandran & Rogers-Ramachandran, 1996; Feinberg, 2011; Subedi & Grossberg, 2011). Overall, computational psychiatry can be utilized to deliver explanations and offer targets for finding possible treatments or strategies for coping with symptoms of various disorders.

Using a computational approach to describe complex cognitive phenomena is not entirely new. Already in 1982, computational approaches to investigating visual processes have been described in David Marr's *Vision* (Marr, 1982). His general framework has sparked debates about levels of explanation, the nature of computation, externalism vs. internalism in computational theories of mind, association of computation and content, as well as top-down versus bottom-up methodology (Marr, 1982; Shagrir, 2010). Marr argues that a computational theory consists of two building blocks that are the "what", the "why". Concretely, the "what" describes the function that is calculated, while the "how" specifies the algorithm of use, to calculate the function (Marr, 1982; Shagrir, 2010). In his framework that fundamentally influenced the philosophy of neuroscience but was, however, not without critics, he postulates three levels of information processing. Specifically, he loosely couples three different aspects that are "Computational theory", "Representation and hardware", and "Hardware implementation". His description of the three levels is presented in Table 1. Bayesian inference of the latent causes of sensory evidence as a goal can thereby be seen as the first level of Marr (computational theory) (Marr, 1982; Aitchison & Lengyel, 2017). The predictive coding framework can be seen as one biologically plausible mechanism that reflects the second level (representational/ algorithmic implementation) as a utilization of Bayesian inference (Heinz et al., 2019; Aitchison & Lengyel, 2017; Valton, Romaniuk, Steele, Lawrie, & Seriès, 2017). The third level (hardware implementation) entails that every algorithm must be executed within a physical system. This physical system can be of biological origin (e.g. humans) but importantly can also be an artificial agent (e.g. robots). Particularly, in the case of schizophrenia in humans, the physical system often entails neurobiological and neurophysiological anomalies, e.g. structural deviations and neurotransmitter irregularities.

3.2. Computational modeling of self-disorders in schizophrenia

3.2.1. Explaining the self-disorders in schizophrenia

Previous attempts at explaining the self-disorders in schizophrenia investigated them on the symptom level, with the goal of trying to explain symptoms rather than viewing them as the root cause of the mental disorder itself. Specifically, current diagnosis schemas for schizophrenia often involve the diagnostic tools "Structured Clinical Interview for DSM Disorders" (SCID), "the Scale for the Assessment of Positive Symptoms" (SAPS), and the "Scale for the Assessment of Negative Symptoms" (SANS), that assess schizophrenia on positive, negative, and cognitive symptoms. However, newer research in schizophrenia points to the notion that schizophrenia rather represents a fundamental self-disturbance (or "ipseity-disorder") out of which the positive, negative, and cognitive symptoms emerge. Nonetheless, these self-disturbances are often neglected in contemporary psychiatry. Symptoms include feelings like a long-time persisting identity void and feeling of self-transformation. Other neglected symptoms are a disturbed stream of consciousness, self-awareness, corporeality, demarcation, and existential reorientation, all of which are interrelated. The Research Domain Criteria (RDoC) can provide a more modern taxonomical framework that address the heterogeneity of mental disorders (Insel & Lieberman, 2013). Furthermore, schizophrenia can be assessed in more phenomenological detail using the "Examination of Anomalous Self-Experience" (EASE) instrument (Parnas et al., 2005), a semi-structured clinical interview focusing more strongly on the experiential and phenomenological anomalies of schizophrenia spectrum disorders. Different explanatory approaches have been used, namely, machine learning approaches, comparator model-based approaches, biological approaches, predictive coding-based approaches, as well as circular inference-based approaches (Section 3.2.3). In this section we will present example studies that were based on these approaches.

Machine learning approaches have been used to identify variables that have predictive value for the clinical prognosis of patients with schizophrenia. Specifically, Koutsouleris et al. (2016) used Kaplan–Meier log-rank analyses to predict discontinuation and readmission to the hospital in patients with poor versus good treatment outcome prediction. In addition, generalized linear mixed-effects models were used to identify factors that predict a positive or negative treatment outcome of patients with schizophrenia after 4 and 52 weeks. Concretely, previous depressive episodes, male sex, and suicidality were all identified as risk factors for the one year period. Additionally, unemployment, poor education, functional deficits, and unmet psychosocial needs predicted a bad outcome for both the 4-week and 1 year outcomes. Furthermore, their analysis was used to asses the efficacy of and comparison between certain antipsychotic medications (Koutsouleris et al., 2016).

Beside general prognostic tools, many models that focus on explaining self-disorders have been proposed. One of the earliest was the comparator model (Feinberg, 1978; Wolpert, Ghahramani, & Jordan, 1995) (see Section 2.3). With the help of the comparator model, studies aimed at explaining self-disorders such as in patients suffering from schizophrenia and reporting thought insertion originating from external sources. According to comparator model-based approaches, if the efferent copy is not correctly transmitted, this causes a mismatch between the representation or prediction of a movement vs. the movement that was actually executed. However, sometimes patients with schizophrenia feel more agency, and also the role of dopamine in the model is lacking. Specifically, the neurotransmitter dopamine is involved in the encoding of salience (precision) during the processing of information (Heinz et al., 2019). Since this crucial part of encoding the salience of a stimulus is lacking, the comparator model can be seen as reductionist and unable to explain certain symptoms (e.g. thought insertion) (Frith, 2012).

The investigation of aberrant glutamate signaling between brain regions has also been a focus of computational psychiatry studies.

For example, in one study, synaptic disinhibition on interneurons through NMDA receptor disturbances has been modelled as a network model of spatial working memory using behavioral data from participants who performed a spatial working memory task. Some of these participants have been given ketamine, thus inducing NMDA disinhibition and simulated certain symptoms of schizophrenia (Murray et al., 2014).

3.2.2. Predictive processing-based models

Studying disorders in perception of the bodily self have been recognized as promising avenues for modeling mechanisms hypothesized to underlie symptoms of schizophrenia. In a comprehensive review, Lanillos et al. (2020a) reviewed and analyzed neural network models of autism spectrum disorders and schizophrenia, and compared Bayesian approaches such as circular inference and predictive coding models. They discussed neural network model implementations that resulted in similar phenomena observed in autism spectrum disorder and schizophrenia based on underlying mechanisms such as circular belief propagation (Section 3.2.3), weak priors and aberrant precision weighting which leads to an imbalance in integration of priors and sensory signals (Philippsen & Nagai, 2020a; Philippsen & Nagai, 2020b), and network dysconnectivity leading to motor phenomena similar to those observed in schizophrenia (Yamashita & Tani, 2012).

In predictive coding, perception is the result of a balanced integration of priors with sensory input. This balance is disturbed if the precision of the prior is higher (hyper prior), leading to a stronger reliance of the posterior on the prior, and less on the sensory evidence. Low prior precision would then lead to a stronger reliance on sensory evidence. This has been implemented in a series of studies with computational neural network models (Philippsen & Nagai, 2020a; Philippsen & Nagai, 2020b), that showed how aberrant reliance on priors during training, could result in impairments in the network's internal representation, as well as replicated behavioral findings from humans and chimpanzees in a representational drawing task. These results are discussed in the context of autism spectrum disorder, as the heterogeneity in symptoms could be explained by this demonstrated mechanism of over-reliance on either predictions or sensory evidence in the course of development (Philippsen & Nagai, 2020a).

As displayed in Fig. 3, the formation of delusion can be explained as a failure to integrate sensory experience with one's prediction (Fletcher & Frith, 2009; Griffiths, Langdon, Le Pelley, & Coltheart, 2014; Garety & Hemsley, 1997; Gray et al., 1991), while the delusions are maintained by the disproportionate prediction error that iteratively reconsolidates and thus strengthens the delusions, despite contradicting evidence (Corlett, Krystal, Taylor, & Fletcher, 2009).

On a purely mathematical sense, and as displayed in Fig. 3, in the case of higher precision of the prior, the perception can be computationally more influenced by the prior, and there will be less reliance on the sensory evidence. In the same way, in the case of lower precision of the prior, the integration of the perception will rely more on the sensory evidence, and the posterior will shift towards it, i.e., there is a trade-off between the precision of the prior, and the variance of the sensory evidence.

Nowadays, a strong emphasis on explaining self-disorders computationally relies on the predictive coding framework. However, the classical interpretation of the predictive coding framework struggles to explain symptoms that are detached from kinematics and sensations (e.g. thought insertion) (Frith, 2012). Specifically, it is debated how predictions can be made and how the prior is updated when sensory feedback of internal processes (e.g. thoughts) is lacking. This is because it is challenging to view thoughts as actions in the context of the generative predictive model (forward model) (see Section 3.3 for discussion).

3.2.3. Circular inference

Circular inference or circular belief propagation is a promising computational approach that builds on Bayesian modeling and the predictive coding framework (Jardri & Deneve, 2013). In order to make sense of the world, the brain uses a generative model that hierarchically represents the causal links between variables that underlie events. This model consists of nodes that represent these (latent) variables, and edges that represent conditional dependencies between them. Bottom-up processing is the process of sensory information going up the hierarchy in a feedforward way. At the same time, top-down processing refers to prior information passing down the hierarchy as feedback.

The sum-product algorithm or belief propagation is one way to perform inference in this generative model. Information is propagated through the system in a feedforward and feedback manner, in the form of beliefs. These beliefs regarding the underlying variables (or causes) are calculated at each node, which sends messages to neighboring nodes. Each node then sums up the information from all its neighbors. The messages passing from node to node depend on the belief of the sending node, *after subtracting* the effect that the receiving node has on the sending node. It is important to subtract the message from the receiving node, because otherwise, the algorithm would produce loops, i.e. bottom-up and/or top-down information would reverberate. In such a situation when there are loops of bottom-up or top-down information, causes are treated as effects and vice versa. The brain continuously integrates top-down and bottom-up information of different entangled feedback loops. If an effective control mechanism is lacking, processed top-down information might be reused as sensory evidence and thus over-counts the actual sensory evidence. This same process could also hold true for bottom-up processes; By reverberating sensory information to a lower hierarchical level, this information can be mistakenly used as top-down expectation, leading to multiple accounting of "old" priors and sensory information. This might ultimately lead to aberrant belief formations, i.e., prior beliefs that are echoed to lower level hierarchies might be misinterpreted as sensory evidence which could explain the "jumping-to-conclusions" bias of patients with schizophrenia in a sense of an over-representation of weak sensory evidence (Huq, Garety, & Hemsley, 1988).

In this framework, bottom-up evidence and top-down predictions are echoed back, leading to a repeated use of already processed information which can explain hallucinations and delusions. Using the same information multiple times is usually avoided when every excitatory loop is compensated by an equally strong inhibitory loop that predicts and cancels out informational redundancies. However, if that is not the case (e.g. in schizophrenia) circular inferences might occur.

In its simplest form, the model consists of three nodes that pass messages up and down the stream. Redundant information is created by sending both information upwards and downwards the information stream which is normally subtracted to avoid redundancies. However, according to this model, some information is not removed in a neuronal system of patients suffering under certain clinical conditions, such as schizophrenia (Jardri & Deneve, 2013; Leptourgos et al., 2017). Fig. 2 presents the model and examples of behavioral consequences of circular inference. We used the experience of paranoia of being followed by the CIA or any other secret service as an example, because it is very common in patients suffering under schizophrenia. In the case of climbing circular inferences (stronger likelihood, weaker priors), sensory evidence is reused, thus the person *expects what they see*. Even weak evidence for a siren (e.g. reinterpretation of a similar sound), will result in an expectation of a siren, reinforcing itself. In the case of descending circular inferences (stronger priors, weaker likelihood), the person *sees what they expect* (e.g. expecting to be in danger results in hearing sirens). The case of both ascending and descending reverberating loops, results in the formation of a "frustrated network". In such a network, mutually exclusive facts might be experienced at the same time. For example, hearing a siren and not hearing a siren from the same car. The strength of the belief update is proportional to the prediction error, weighted by the precision ratio of the inverse variance of the likelihood and priors (Petzschner, Weber, Gard, & Stephan, 2017).

In sum, the repeated use of sensory evidence as higher order top-down expectation leads to a sensory over-representation which could explain the origin of hallucinations, delusions, and the subsequent consolidation of delusional beliefs, which is common in patients with schizophrenia. An example of this is the common experience of patients with schizophrenia of being followed by a secret intelligence service. Even little evidence for a siren (e.g. a similar sound) could be reused multiple times which results in hearing a siren and the formation of the belief that one is followed by a secret intelligence service and imminent danger. At the same time, the feeling of being surveilled could enhance the perception of seeing black cars everywhere (that are often used by secret intelligence services) which leads to hallucinating a siren, further enhancing the paranoia.

Some patients are also impaired in their downward loops, and thus, over-count their priors which could explain the diversity in patients' behavior (Jardri & Deneve, 2013; Tandon, Nasrallah, & Keshavan, 2009). Specifically, some patients with schizophrenia sometimes assume to have significantly more agency over their own actions, while other patients experience significantly less agency compared to their objective level of agency (Proust, 2006). The circular belief propagation model aligns well with dysconnectivity hypotheses of imaging studies in patients in schizophrenia (Jardri & Deneve, 2013; Notredame, Pins, Deneve, & Jardri, 2014; Liang

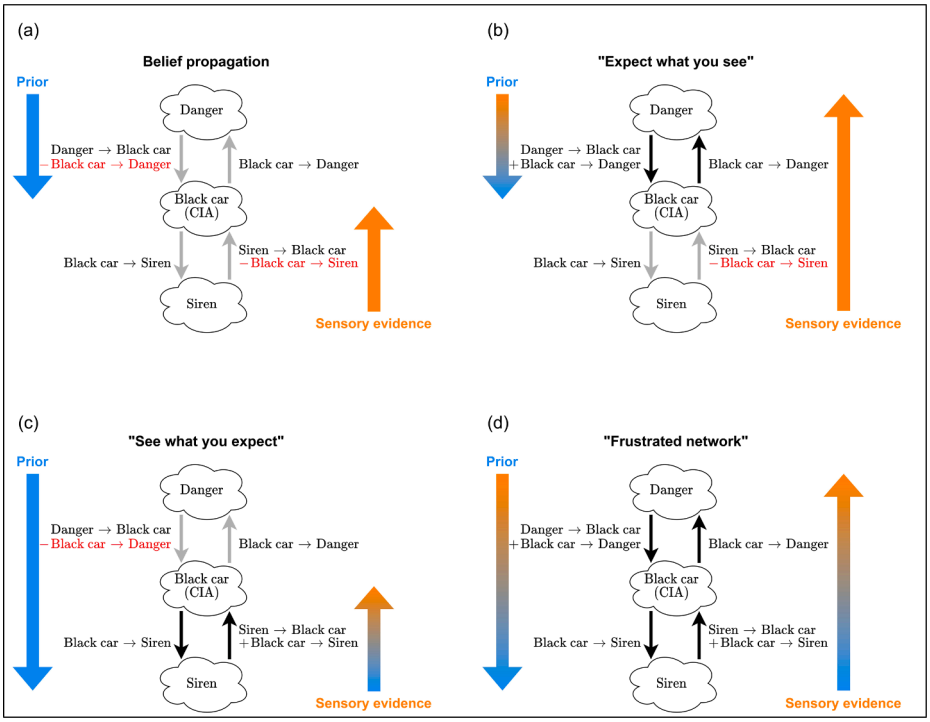

Fig. 2. Belief propagation and circular inference. (Jardri & Deneve, 2013; Leptourgos et al., 2017). (a) The model consists of three nodes that pass messages up and down the stream (prior in blue, sensory evidence in orange). Redundant top-down and bottom-up information is normally subtracted (in red, subtraction is indicated by the "−" sign) to avoid reverberation. Consequences of circular inference: (b) In the case of climbing circular inferences, redundant sensory evidence is reused (in black, reuse is indicated by the "+" sign), reverberating to corrupt the prior, creating the effect of *"expect what you see"*. (c) In the case of descending circular inferences, the prior is reverberated through the system, corrupting sensory evidence, creating the effect of *"see what you expect"*. (d) Here, reverberating redundant information from both ascending and descending loops results in the formation of a "frustrated network". In such a network, mutually exclusive facts might be experienced at the same time. (Adapted from Leptourgos et al., 2017). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2006; Amad et al., 2014). Evidence for that comes from e.g. the latent inhibition paradigm (Lubow & Moore, 1959; Lubow, 1973; Swerdlow, Braff, Hartston, Perry, & Geyer, 1996) that tests the capability to filter irrelevant stimuli, a capability that is impaired in patients with schizophrenia. This is an expected outcome when upward inhibitory loops are impaired, leading to over-counting of unreliable sensory information (Jardri & Deneve, 2013).

Neurobiologically, the predictions are likely transmitted via descending connections of N-methyl-D-aspartate (NMDA) receptors, while prediction errors are more likely to be transmitted via AMPA-receptors, and maybe supported by NMDA connections at ascending connections. The precision weighting is gauged via the postsynaptic gain, reflected by Dopamine, Achetylcholine, and GABA (Bastos et al., 2012; Petzschner et al., 2017; Corlett, Taylor, Wang, Fletcher, & Krystal, 2010; Friston & Kiebel, 2009; Corlett, Honey, Krystal, & Fletcher, 2011; McCormick, Wang, & Huguenard, 1993; Adams, Stephan, Brown, Frith, & Friston, 2013).

Belief propagation is considered a biologically plausible mechanism because the algorithm is comparable to mechanisms of propagation and integration of neuronal activity in neural microcircuits (Leptourgos et al., 2017). Specifically, inhibitory loops can remove information redundancies either by reducing the feedforward information of the top-down flow, or by diminishing bottom-up information. Belief propagation can be implemented on the neuronal level by balancing excitation and inhibition. Pathways in the human visual cortex show different frequency band synchronicity (microcircuits), influenced by feedforward and feedback projections (Michalareas et al., 2016). These findings provide support for the possible implementation of belief propagation on the neuronal level, since imbalances in excitation and inhibition in these microcircuits could lead to reverberations, leading to circular inference.

Imbalances in excitation and inhibition could occur in local and global inhibitory loops. In fact, inhibitory deficits are present in patients with schizophrenia. These have been correlated with GABA deficits and have been shown to predict the formation of an aberrant belief system (Denève & Jardri, 2016). These findings are also in line with brain imaging findings within the same population (Jardri et al., 2016). Looking at the long-range inhibitory loop, the thalamic and limbic loops have been identified to be involved in neocortical inhibition (Maffei, 2017). Of special interest are the neocortical-striatal pathways, that are involved in hallucinations and more broadly, in psychosis (Howes et al., 2011; Rolland et al., 2015), and could be related to inhibition of feedback signals.

The inhibition of feedforward signals could be associated with thalamocortical inhibitory loops. Dysfunctions in long-range inhibition could be due to a dysconnectivity between the thalamus and the visual cortex, which has been found in patients with schizophrenia (Yang et al., 2014). The thalamus constitutes a junction for sending and receiving information to and from the cortex, therefore it could be implicated in inhibition of feedforward signals (Rolland et al., 2015). In patients with schizophrenia, the effective connectivity between the thalamus and the visual cortex is reduced, leading to a disruption of causal information flow (Iwabuchi & Palaniyappan, 2017).

3.3. Open Questions and challenges in modeling self-disorders in schizophrenia

While the comparator model has had success in explaining certain symptoms of schizophrenia, including certain aspects of the self-disorders (Frith, 2012), some open questions still remain. It is still unclear why specific symptoms in patients with schizophrenia often differ markedly from one individual to another. The comparator model does not explain inter-individual differences such as elevated sense of agency in some patients, and reduced sense of agency in others. Also, this model does not take into account the elevated role of dopamine that is involved in the precision weighting in patients with schizophrenia (Frith, 2012). In addition, the model struggles to explain thought insertion, a first-order symptom experienced by half of all patients with schizophrenia. The most important question is whether thoughts can be treated the same way as actions, since thoughts are not linked with intentions (Sterzer et al., 2016). The argument is that in order to have the intention to think, this needs to be preceded by an intention, leading to an endless loop (Akins & Dennett, 1986; Stephens & Graham, 2000; Heinz, 2014). Others question whether a comparator is at all a valid model to explain thought patterns (Gallagher, 2004), since it is counter-intuitive that thinking needs to "explain away" sensory signals. Furthermore, one has access only to self-generated thoughts, so usually a distinction from others' thoughts is not necessary. Additionally, the comparator model encounters problems in explaining the difference between thoughts transferred by another individual and intrusive, unwanted thoughts in general (Sterzer et al., 2016; Stephens & Graham, 2000; Heinz, 2014; Gallagher, 2004; Vosgerau & Newen, 2007; Brewin, Gregory, Lipton, & Burgess, 2010).

However, there are some concerns about the predictive coding framework. A major concern comes from the methodological point of view. As put forward by Popper (1959), a theory like predictive coding must be proven useful in predicting the results of the observation. Inferring the values of the parameters by using the observation refers to an "inverse problem" and is invalid, since observations should only be used to falsify possible solutions (Tarantola, 2006). This is however not the case in predictive coding. Another criticism of the predictive coding approach is that it does not take the enactive, embodied, and encultured aspects of phenomenology into account (Humpston & Broome, 2020; Allen & Friston, 2018). Specifically the enactive approach emphasizes the social, cultural, and relational aspects of the illness which is mostly neglected by current predictive coding models (Kiverstein, 2020).

Another critique of predictive coding is that there is no agreement among the scientific community, if predictive coding sufficiently explains thought insertion. Some argue its explanatory value only applies to action and perception and not to thoughts, questioning whether we can treat a thought like an action (Frith, 2012). Arguing in favor of a predictive coding account of thought insertion, Sterzer et al. (2016) assert that inserted thoughts are interpreted as surprising, and in no logical continuity with prior thoughts and are thus interpreted as "coming from nowhere" and inserted by another agent. They argue that thoughts constitute prior beliefs about what thought is likely to arise next which fits the predictive coding framework. Their second argument is, in contrast to Frith (2012) that thoughts can be just like perceptions and actions reduced in sensory prediction and increased in sensory precision, leading to a higher prediction error and thus surprising thoughts. In other words, the salience of thoughts is the basis for the perception of thought insertion and it is rather the high salience of the thought caused by imprecise prior beliefs, rather than the content itself (or disturbed

interoception as in the comparator model (Campbell, 1999)), thus providing evidence for the predictive coding approach. Benrimoh, Parr, Vincent, Adams, and Friston (2018) noted that the increase in sensory precision may lead to the realistic and "loud" quality of hallucinations. Kaminski, Sterzer, and Mishara (2019) give a case report in which a patient suffering from schizophrenia describes "seeing rain" that demarcates him from his surroundings. They conclude that self-disturbances can be seen as adaptive coping-strategies to compensate for the rupture in the perception–action cycle in patients with schizophrenia. Interestingly, many patients report uncertainty if they are asked to indicate if the voices they are hearing are auditory hallucinations, thought phenomena, or something in between, revealing there is an underlying spectrum of hearing voices and interpreting them as thoughts (Humpston & Broome, 2016). Overall, the phenomenology of psychotic experience can be explained well by the dynamics of hierarchical predictive coding since both utilize the perception–action cycle within the hierarchical formalism (Kaminski et al., 2019).

However, on a computational level, the posterior perception can either be shifted through the precision of the prior belief, or the precision of its likelihood. This makes it unclear whether patients with schizophrenia have stronger, or weaker priors. There is indication that in general, priors are weaker for delusions, and stronger for hallucinations (Corlett et al., 2019; Stuke, Weilnhammer, Sterzer, & Schmack, 2019). The formation of imprecise prior beliefs either occurs by a faulty acquisition of prior beliefs, or through an inability to use prior beliefs as inferences (Heinz et al., 2019). Faulty prior beliefs can be consolidated either by misleading sensory information (e.g. being followed by black cars from the secret service), or if the encoding and/or detection of sensory information is disturbed (Heinz et al., 2019; Lisman & Grace, 2005). Another possibility for faulty prior beliefs is if the detection and/or computation of higher order prediction errors are perturbed or inadequately precise (Heinz et al., 2019; Adams et al., 2013; Adams, Huys, & Roiser, 2016). The inability to use prior beliefs as inferences on the other hand, could occur if higher order beliefs contain erroneous predictions of the volatility of new sensory stimuli (e.g. mistake a tree branch for a dangerous snake in the park) (Heinz et al., 2019). These different possibilities and individual differences of impairment could explain the heterogeneity in symptoms of schizophrenia (Sterzer, Voss, Schlagenhauf, & Heinz, 2019). For a simplified graphical overview on the changes of priors, posteriors, and likelihood in patients with schizophrenia compared to neurotypical persons, see Fig. 3. Notably, this simplified model is a moment-to-moment depiction of a highly dynamic process with adaptation throughout, while the relationship between delusions and prior and sensory evidence usage is more complex (Stuke et al., 2019; Schmack et al., 2013; Teufel et al., 2015; Alderson-Day et al., 2017; Powers, Mathys, & Corlett, 2017). Contrarily, there is evidence that higher order priors have a higher impact on delusions, while lower-order priors are reduced (Stuke et al., 2019; Schmack et al., 2013; Schmack, Rothkirch, Priller, & Sterzer, 2017). More concretely, while in the formation of delusions the prior is decreased, higher order priors may be increased to compensate for the lower-level priors.

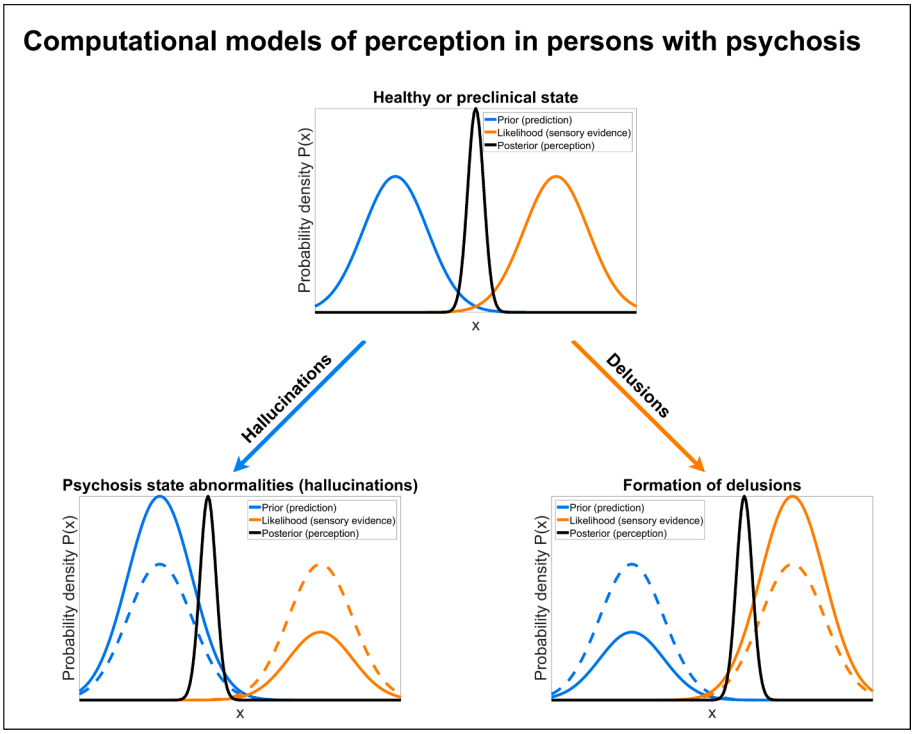

Fig. 3. The computational foundations of perception in psychosis. Top: In both neurotypical states and in preclinical states of psychosis, the computational mechanisms are undisturbed. Bottom left: In hallucinations (state abnormalities), either the prior precision is increased, the likelihood precision is decreased, or both (indicated by the difference between dashed and solid line of the respective color. The precision is represented as the alpha of the Gaussian curve). This shapes the perception (black curve) towards an over-reliance on top-down predictions. Bottom right: In the formation of delusions (trait abnormalities), either the prior precision is decreased, the likelihood precision is increased, or both. This shapes the perception towards an over-reliance on bottom-up sensory evidence.

Subsequently, the low-level prior suppression that initially leads to the formation of delusions is strengthened and maintained by higher order priors that maintain and strengthens the delusion (Stuke et al., 2019).

The circular inference model can explain the origin of hallucinations, delusions, and the consolidation thereof. Additionally, possibilities of disruptions in the feedforward, and feedback loops can explain the diversity of symptoms that can be on either side of the extremes. Another advantage is the biological plausibility of this theory that falls in line with the dysconnectivity hypothesis of schizophrenia (Jardri & Deneve, 2013; Notredame et al., 2014; Liang et al., 2006; Amad et al., 2014). However, there is still a need to validate the framework on the neurophysiological level.

4. Cognitive developmental robotics (CDR)

4.1. Guiding principles of CDR

Cognitive robotics is concerned with providing robots with a cognitive architecture that will allow them to have complex interactions in a complex world, and to be able to adapt to changes in the environment. A branch of this field of research, developmental robotics, focuses on the processes and mechanisms that allow lifelong and open-ended learning of new knowledge and skills in an embodied system, modeled after human infant development.

An important design principle in developmental robotics is modeled after the human development process i.e. as a process of incremental acquisition of knowledge from experience, in a physically embodied system (Lungarella et al., 2003; Asada et al., 2009; Asada, MacDorman, Ishiguro, & Kuniyoshi, 2001; Stoytchev, 2009; Cangelosi & Schlesinger, 2015). The focus here is on identifying and implementing those basic behavioral and computational building blocks that enable the autonomous bootstrapping of motor and cognitive skills in artificial agents. Developmental robotics can therefore provide new understanding regarding the emergence of higher cognitive functions in humans.

Different issues are faced by researchers in this field. Incremental learning is a challenge yet to be solved in the artificial intelligence community (Nguyen et al., 2020). Updating an artificial neural network in an online fashion typically deteriorates the knowledge that has been previously acquired. Current measures employed to balance the stability and plasticity of these models still barely resemble those of the human brain.

Another issue regards the problem of scalability. This arises when hidden assumptions in pre-programming make it increasingly difficult for the robot to autonomously adapt to situations that violate these assumptions. From this emerges the principle of *verification* in cognitive developmental robotics (Stoytchev, 2009). Verification requires that the robot will be in charge of testing and verifying everything that it learns. Furthermore, for a robot to be able to verify what it learns, then we need the robot to act upon the world, and for this the robot needs to have a body (Stoytchev, 2009).

4.2. Modeling the self in CDR

While cognitive models of self-recognition in biological agents used to be very different from self-recognition models in artificial agents, novel algorithms and techniques of machine intelligence are developing that are coming closer to biological processes i.e. timing, spatial information, and sensorimotor contingencies, and even hardware (e.g. neuromorphic computing: Schuman et al., 2017), ultimately simulating mechanisms thought to underlie the self in humans more thoroughly (Lanillos et al., 2020b; Stoytchev, 2009; Nguyen et al., 2020; Lanillos, Dean-Leon, & Cheng, 2016; Gold & Scassellati, 2009). In the following sections we will review some of the modeling approaches used in CDR.

4.2.1. Comparator models in CDR

Comparator models are widely used in cognitive robotics. Optimal control theories (Wolpert & Kawato, 1998) identify two types of internal models: the inverse model maps a desired sensory state to the motor action that will most likely achieve it, and the forward model maps a motor action to a sensory outcome.

The inverse model is used for action selection and the forward model is used for mapping action-effect pairs and to compare the achieved sensory state to the predicted one. The sensory prediction is based on the motor command (the efferent signal) and is then compared to the afferent sensory signals. According to the comparator model theory of the sense of agency, the congruence between predicted and observed sensory consequences of an action will give rise to the sense of agency (Gallagher, 2000).

A study by Lang, Schillaci, and Hafner (2018) investigated learning in a humanoid robot that learned the visual sensory outcomes of self-generated movements through a self-exploration behavior. The sensorimotor experience acquired through this self-exploration was used as training data for a deep neural network integrating convolutional layers. The deep forward model mapped proprioceptive (e.g. initial arm joint positions) and motor data (motor commands) onto the visual outcomes of these actions. This forward model was then used in two experiments. First, the forward model generated visual predictions of self-generated movements, which were then compared to actual visual outcomes to compute a prediction error. Higher prediction errors occurred when an external subject was performing actions in front of the robot, in contrast to when the robot was only observing itself performing the same arm movements (Lang et al., 2018). This is in line with the notion that prediction errors can be utilized for self-other distinction by way of body ownership, which is thought to be linked to the sense of agency (Braun et al., 2018; Ma & Hommel, 2015b; Ma & Hommel, 2015a; Möller, Braun, Thöne, Herrmann, & Philipsen, 2020). The results showed that prediction can be used to attenuate self-generated movements, and to create enhanced visual perceptions. The sight of objects was still maintained even when the view on the object was occluded by the robot's arm movement (Lang et al., 2018). This indicates that similar processes could be used to further our

understanding of the sense of object permanence and short term memory systems in humans.

In a biologically inspired model, proposed in Schillaci, Ritter, Hafner, and Lara (2016), multimodal body representations were acquired through learning and predicting the robot's ego-noise i.e. auditory noise from the motors during movements. In an ego-noise attenuation experiment, a predictive process was implemented by a forward model, that took as inputs coherent and incoherent proprioceptive and motor information. The effects of the coherent and incoherent information were shown in the performance of the ego-noise suppression. Ego-noise attenuation was found to be stronger when the robot was the owner of the action. Ego-noise attenuation was less pronounced when the robot was only listening to the noise of a simulated moving robot. This is because greater prediction errors occurred when motor and proprioceptive information was incongruent with the predicted ego-noise. The "surprise" caused by this incongruence allowed the artificial agent to classify self-generated actions and those generated by other subjects differently.

Utilizing the fact that self-produced signals are likely to be maximally predictable, Schillaci and colleagues developed an artificial system that used the prediction of self-generated auditory information for sensory attenuation (Pico, Schillaci, Hafner, & Lara, 2016; Schillaci et al., 2016; Bechtle, Schillaci, & Hafner, 2016). This allowed the robot to classify self- and other- generated auditory signals. Concretely, the robot identifies external auditory information as more salient, since the predictions were matched with the self-generated signals. When there is a match of predictions and sensory information, the comparator model filters out self-produced signals, leading to a stronger salience (or surprise) of external signals.

However, the explanatory value of the comparator model for the sense of agency is debated. Zaadnoordijk et al. (2019) argue that the sense of agency requires not only the representation of the match between predicted and observed sensory consequence of an action (the prediction error), but also a representation of an action performed by the agent (an "ownership predicate"), and a representation of the causal relation between the (own) action and its effect (Zaadnoordijk et al., 2019).

Lanillos et al. (2020b) used a "double comparator" model for robot body estimation and self-recognition and self/other distinction tasks. In this model the first comparator is used when the robot needs to infer the most plausible location of its arm (the learned forward model) by using the prediction error between observed and predicted sensory input. The second comparator considers the spatiotemporal contingencies between visual input from optical flow and motor actions (joint velocity) of the robot to compute the probability of the sensor values being generated by itself.

4.2.2. *Active inference models*

Predictive coding questions the need for an inverse model and a resulting efferent copy for the achievement of goals. In fact, optimal control theories present difficult issues to solve, among which the ill-posed problem of learning such inverse models (Pickering & Clark, 2014; Dogge, Custers, & Aarts, 2019). In predictive coding there are no reward or cost functions to optimize behaviors, instead, these are replaced by priors about sensory states and their transitions (Friston, Samothrakis, & Montague, 2012).

An overview of the differences between internal models in optimal control theory and in predictive coding can be seen in Fig. 4 and Fig. 5. Fig. 4 depicts the classical approach, in which an inverse model provides an efference copy of the motor command to an auxiliary forward model. In the integral forward model à la predictive coding, motor commands are replaced by top-down proprioceptive predictions (see Fig. 5). These can be viewed as control states that are translated into muscle-based coordinates fulfilled by classical reflex arcs (Friston et al., 2012).

Under the Bayesian brain hypothesis, the brain is an inference machine that can make sense of the world based on partial information. According to active inference, empirical information is adapted to the world model either by changing the belief, or by performing an action that would alter the world according to predictions. Active inference models in cognitive robotics usually entail a

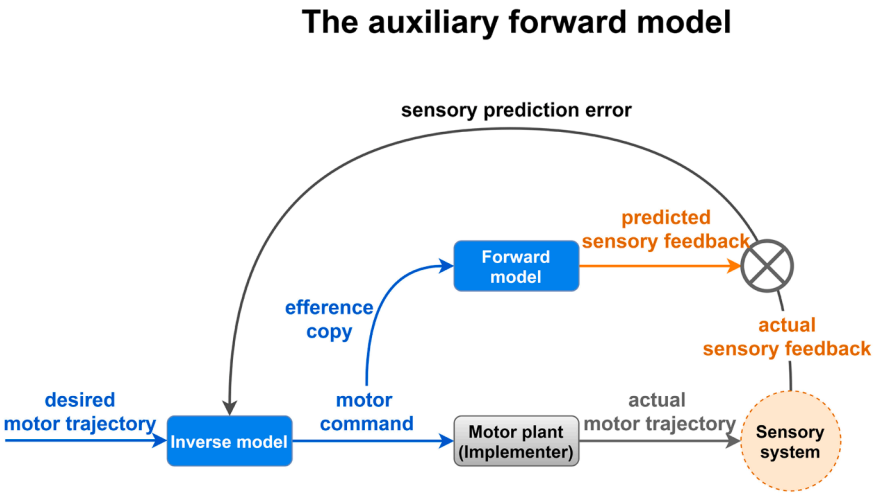

Fig. 4. The auxiliary forward model. In the auxiliary forward model architecture, the inverse model outputs a motor command, which serves as input to the forward model, that predicts the sensory feedback. (Adapted from Pickering and Clark, 2014).

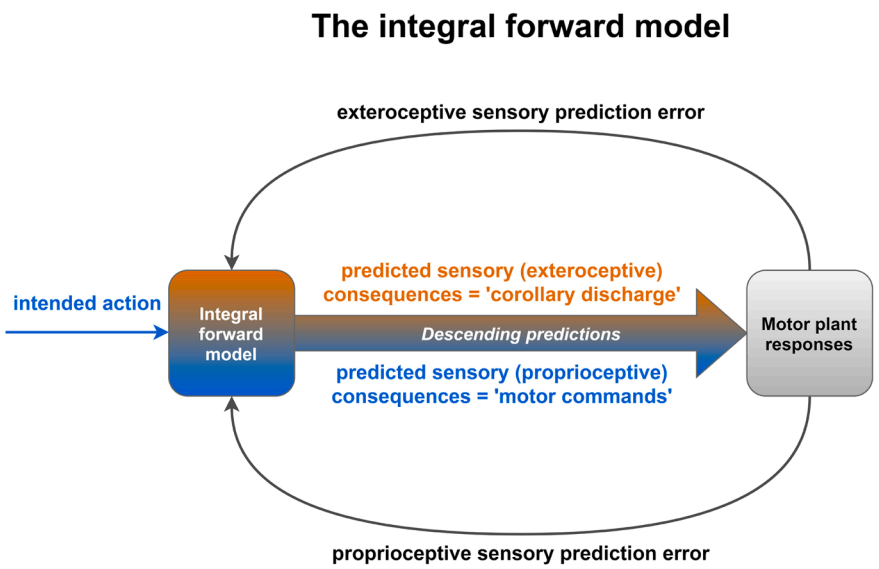

Fig. 5. The integral forward model. In the integral forward model architecture, no efference copy is required. Instead, predictions from the forward model are used as action commands. (Adapted from Pickering and Clark, 2014).

generative model that makes predictions, minimizing prediction error between expected sensory effect of an action and observed sensory effect, thereby minimizing the free energy. Minimizing free energy, prediction error, or "surprise" can be achieved by either adapting the (generative) model that makes the predictions (perceptual inference), or acting on the world, thus changing the sensory information (active inference). Active inference is implemented when actions are selected such that the free energy will be minimized. For example, Tani and White (2020) reviewed a series of studies that employed analogous models for minimizing the free energy.

4.3. Robotic modeling of self-disorders in schizophrenia

It has been hypothesized that schizophrenia symptoms, and the disorder itself stem from system-level dysconnectivity (Friston & Frith, 1995; Stephan, Baldeweg, & Friston, 2006). Specifically, aberrant prediction-error signals stemming from underconnected neural networks, would change the goal-orientation in the network, even without an overt change in behavior. Yamashita and Tani (2012) simulated the network functional dysconnectivity in a neural network-driven model in a humanoid robot. They used a hierarchical model of top-down and bottom-up network representing the intention/goal and the lower sensorimotor level, respectively. To test network dysconnectivity, they slightly modified the connective weights between the higher and the lower levels of the model by adding varying levels of random noise. These represent changes in synaptic connectivity in the brain that are thought to occur in patients with schizophrenia. The task for the robot was to repeatedly move an object in two different ways, depending on the location of the object. The position of the object was changed by the experimenter at unpredictable times, which produced a temporary increase in prediction error. In turn, this increase in prediction error produced a modulation in the robot's intention state to minimize the prediction error, causing a flexible switching of behavior. Different levels of network dysconnectivity have been manipulated to observe how the robot will deal with a surprising event on both the computational and behavioral level, by observing the spike in prediction error, switching between intentional states, and observed patterns of motor behavior. The findings showed that mild network dysconnectivity produced an increase in prediction error but outwardly normal behavior. However, higher, more severe levels of network dysconnectivity produced spikes in prediction error and irregular switching in intention states, and overt behavioral deficits such as disorganized actions, cataleptic (stopping or freezing) or stereotypic (repetitive) behavior. These findings were consistent with similar phenomena observed in patients with schizophrenia (Yamashita & Tani, 2012).

Artificial systems can also enter delusional states, as a result of dysfunctions in predictive learning relating to the forward model. Predicting outcomes of own actions on the body and on the environment is a crucial part of predictive learning in both biological and artificial systems. This requires the system to disambiguate self-induced sensory input from externally generated sensory information, using a forward model to essentially filter the sensory signal as to attenuate the reafferent component and learn from the residual signal. When sensory information is too strongly filtered, over-reliance on priors might cause a "delusional loop", in which the forward model is overly weighted, and learning is stalled (Kneissler, Drugowitsch, Friston, & Butz, 2015). In this case learning will not reach proper convergence because the uncertainty of the forward model is not addressed. Specifically, if the forward prediction filters the sensory input too strongly, hyper-confidence in the prediction can lead to a delusional state that completely ignores new incoming information, and stalls learning. To disentangle this problem, Kneissler et al. (2015) proposed a Bayes-optimal linear forward model, which they call "Predictive Inference and Adaptive Filtering" (PIAF). This method filters incoming sensory information, but simultaneously improves the forward model, thus preventing delusional states.

When approaching to model self-disorders it is important to incorporate in the architecture the underlying mechanisms thought to

be involved in the emergence and constitution of the self, such as body representations, multimodal integration, and predictive processes (Nguyen et al., 2020). In addition, one should also consider developmental aspects. In simple terms, if we want to study a self-disorder in an artificial agent, we need to provide a plausible model of an embodied emerged self that underwent an iterative and interactive developmental process (Hafner et al., 2020). This is especially necessary when considering self-disorders as resulting from system-level impairments. It follows then, that we also need to consider the measures and metrics for an artificial self (Georgie, Schillaci, & Hafner, 2019; Hafner et al., 2020). With the advantage of being able to look inside the "black box", researchers are able to analyze both computational and behavioral measures and indices (Lanillos et al., 2020b; Hinz, Lanillos, Mueller, & Cheng, 2018) of different aspects of the self.

4.4. Open questions and challenges in modeling the self in CDR

Zaadnoordijk et al. (2019) argue that the match between predicted and observed sensory signals is necessary, but not sufficient to understand the emergence of the sense of agency in humans, nor to bring about the sense of agency in a robot. The argument relies on the question: How does the robot know that the action was produced by itself? there are some possible options: (i) The robot does not know that the action was produced by itself. The categorization of signals to self-generated or non-self-generated comes about as a result of the similarity between the observed signal and the predicted one on the signal level, and the interpretation of these categories as "self" and "other" is done by the researchers. (ii) "self" and "other" are labels that were hard-coded into the robot, and as such, after categorizing the signals (as in (i)), the robot assigns the labels based on its pre-programming. (iii) the match between the observed and predicted sensory signals is used by the robot to infer the cause of that match (here the cause being its own actions), and to infer that the robot is a distinct agent. Option (iii) is the only one which could lead to the sense of agency in a robot, Zaadnoordijk argues. In this view, it is not the match between prediction and observation, nor a comparison between sensory signals, that led to the sense of agency, but the process of inferring the cause of the match. Therefore, in order to gather insight on the emergence of the sense of agency in developmental sciences, and to develop a sense of agency in a robot, one needs to rather focus on the additional inferential process regarding the authorship of the action that caused the observed sensory signal (Zaadnoordijk et al., 2019). Yet, it is not clear whether such an authorship attribution would be a pre- or post-reflective process. An unanswered question in this study is how this attribution could be achieved in an artificial agent. If one assumes that self-perception results from learning sensorimotor contingencies and cause-effect regularities, such an attribution could rely on the quality of the predictions of the learned forward models. A binary self-other labelling depending on this match is perhaps too simplistic, but the same machinery could be used to achieve the additional inferential process presumed by Zaadnoordijk et al. (2019).

In a recent review, Ciria, Schillaci, Pezzulo, Hafner, and Lara (2021) provide a comprehensive review of robotic works employing predictive coding and active inference schemes, highlighting their limitations and suggesting avenues for further research. Ciria et al. (2021) highlighted that different challenges in applying active inference schemes in robotics are still unsolved. Learning is among the most evident ones. Several methods are used and only a few are equivalent to the formulations of the free energy principle. Moreover, learning and testing are often decoupled. Furthermore, little attention has been placed on how multiple modalities—apart from proprioception and vision—can be integrated into a learning mechanism under an active inference scheme. Limited research has addressed the scaling up of the predictive coding paradigm towards higher cognitive capabilities. The question of what are the long-term possibilities of using generative models for perception, action, and planning in cognitive robotics still remains unclear.

5. Discussion

5.1. Crossing computational psychiatry and cognitive developmental robotics

The main advantage of linking computational psychiatry with cognitive developmental robotics is that cognitive and developmental processes in humans can be modelled more thoroughly. Not only is a representation of a body closer to reality, but a body might be an indispensable prerequisite to develop a self, be it biological or artificial (Neisser, 1988; Gallagher, 2006; Varela, Thompson, & Rosch, 2016; Pfeifer & Bongard, 2006). By detecting crucial developmental aspects that lead to the emergence of a self, and analyzing vulnerabilities that might lead to disruptions of self experience, robots allow us to test hypotheses regarding aberrant self-development and functioning in embodied agents. Since increasing evidence points to the notion that self-disorders such as schizophrenia arise from a fundamental disconnectedness from one's body, rather than being a mere epiphenomenal symptom of the disease, only embodied artificial systems would allow investigating the crucial embodiment aspect of these disorders, since cognitive process are deeply entangled with one's body that acts upon the world (for a review, see Wilson, 2002). Therefore, robots can be utilized as a testing ground for computational theories that can also incorporate the enactive approach as opposed to purely computational simulated approaches such as neural networks and AI. Furthermore, by simulating the development of human behavior, cognition, and emotions in artificial agents, we might gain a better understanding of human mental skills, their evolution, as well as how to build more intelligent artificial agents (Weng et al., 2001).

An example for a suitable experiment where computational psychiatry and cognitive developmental robotics could be linked is the "force matching task" (Shergill, Bays, Frith, & Wolpert, 2003; Shergill, Samson, Bays, Frith, & Wolpert, 2005; Bays, Wolpert, & Flanagan, 2005; Bays, Wolpert, Haggard, Rosetti, & Kawato, 2008). According to the comparator model, an agent needs the ability to differentiate one's body and actions from the sensations and events of the environment in order to perceive oneself. To distinguish one's body, the agent additionally needs to integrate multisensory afferent signals. To predict and attenuate the sensory feedback stemming from movement, an agent relies on efferent information, that depend on the forward model processing (Fig. 1) (Kilteni &

Ehrsson, 2017). The brain usually decreases the salience of self-generated sensations (sensory attenuation) compared to externally generated sensations to avoid cognitive overload and to distribute attention to external information that may contain more predictive value (Bays & Wolpert, 2007; Voss, Ingram, Haggard, & Wolpert, 2006; Voss, Ingram, Wolpert, & Haggard, 2008). This underlying principle can for example be seen by the fact that healthy persons usually cannot tickle themselves. The same physical properties of the touch feels less intense when it is caused by oneself, compared to when it is caused by another person, or even a machine (Blakemore et al., 2000; Weiskrantz, Elliott, & Darlington, 1971; Blakemore, Frith, & Wolpert, 1999; Wolpert, Ghahramani, & Flanagan, 2001). This reveals that the sense of ownership is a determining factor in the attenuation of somatosensory information (Kilteni & Ehrsson, 2017). Specifically, the sense of ownership rejuvenates the internal body state representation that in turn sends information to the forward model, generating predictions during voluntary action (Kilteni, Maselli, Kording, & Slater, 2015). Therefore, somatosensory attenuation can also be seen as an indicator for body ownership if the task entails the integration of active movement (Kilteni & Ehrsson, 2017).

Sensory attenuation for self-perception has been studied in robots in the visual (Lang et al., 2018) and auditory (Schillaci et al., 2016) domain, but has received limited exploration in the tactile one. Existing robotics studies using the tactile modality in the development multi-modal body representations and internal models (Gama, Shcherban, Rolf, & Hoffmann, 2020; Lanillos & Cheng, 2018), in fact, do not address the role of prediction and sensory attenuation in self-perception. We encourage further exploration towards this research direction.

Modeling the self and its disorders in computational and robotic systems poses challenges in both computational psychiatry (Section 4.4) and CDR (Section 3.3). We have reviewed some of the relevant challenges in each field. Yet, crossing these two fields might involve further challenges, rooted in both theory and implementation.

Most models that are used in computational psychiatry are probabilistic models that aim to closely represent the neuroanatomical structure of the cortex (e.g. Predictive coding, circular inference, dynamic Bayesian networks, Kalman filters, variational Bayes recurrent neural networks (Friston, 2005; Clark, 2013; Friston, 2012). Indeed, many findings in neuroscience and quantum dynamics (e.g. EPR paradoxon, Bells inequality) point to the notion that the world and the perception thereof is of probabilistic nature (Friston, 2012; Knill & Pouget, 2004; Wetterich, 2020; Mückenheim, 1983; Bell, 1964; d'Espagnat, 1979). Probabilistic and variational aspects of such frameworks can be implemented in robotics and AI, although at the cost of introducing difficulties in their scaling up. Probabilistic modeling of multi-modal integration and incremental learning is still challenging in this field. The probabilistic formalism elegantly explains mechanisms in computational psychiatry. However, several processes can be synthesized in robotics and AI by other means than Bayesian modeling. For instance, precision weighting, here modulating the inverse of the variance of a given distribution, could be modelled through gating systems in deep sensor fusion models.

5.2. *Guiding principles in linking computational psychiatry and robotics, in the study of schizophrenia as a self-disorder*

In Section 3, we described Marr's three levels for understanding complex information processing systems and how to apply them to the study of patients with schizophrenia, and to the development of self-models in robotics. For both, human and robotics studies, the free energy principle can be applied on the first level as a computational theory. On the algorithmic and computational level, the predictive coding framework and active inference can be applied (Tani & White, 2020). On the last level, the hardware implementation, we have the brain in humans in which calculations are made using mainly electricity and neurotransmitter. In patients with schizophrenia, the brain often shows anomalies like grey matter loss, and a dysconnectivity in certain brain areas. In robots on the other hand, the implementation is represented by neural network models within the central processing unit of the robot. Similar to the disconnections in patients with schizophrenia, artificial neural network models in the cognitive architecture of the robot can be disturbed (e.g. by adding noise, see Yamashita & Tani, 2012) to develop robotic lesion studies. Overall, Marr's framework provides a common language for designing comparable experiments between human and robotics cognitive sciences. These implementations can be validated and may ultimately deliver insights to cognitive phenomena like the self and the loss thereof (Tani & White, 2020).

We argue that state abnormalities like hallucinations usually stem from an increase of prior precision, while trait abnormalities like delusions are more likely to stem from a decrease of prior precision (Sterzer et al., 2018). The formation and maintenance of delusions is thereby likely supported by a dysregulated activity of dopaminergic neurons (Heinz et al., 2019). This additional noise leads to a higher "aberrant salience attribution", thus drowning relevant stimuli in noise (Heinz et al., 2019; Heinz, 2002; Miller, 1976; Kapur, 2003). Moreover, this noise-inflation prevents relevant information from gaining enough novelty and salience to be incorporated into one's belief system (Lisman & Grace, 2005; Adams et al., 2016).

5.3. *Deriving novel research topics for robotics from computational psychiatry*

Patients with schizophrenia show consistent impairment in learning tasks that require explicit learning and memory. Implicit processing and learning (especially motor learning) however seems to remain relatively intact (Horan et al., 2008). Illusions can be seen as the difference between the objective and perceived object properties and are products of rational Bayesian evidence that is present in both healthy humans and humans suffering under pathological disorders (Notredame et al., 2014). Evidence points to the notion that patients with schizophrenia are more prone to illusions compared to healthy persons (Notredame et al., 2014).

Besides learning, another interesting aspect of schizophrenia that could be studied with robots are dream states. It has been shown that patients suffering from schizophrenia not only experience sleep problems more frequently, but their dream states seem to be altered as well. Specifically, patients with schizophrenia experience significantly more nightmares compared to healthy controls which is also positively correlated with their subjective distress (Michels et al., 2014).

It has been suggested that the state of acute schizophrenia can be described as a minds' in-between state of waking life and dreaming. While both waking and sleeping states are functional, an in-between state is dysfunctional since the brain attempts to be in two conflicting brain states at the same time (Llewellyn, 2009). Interestingly, many researchers have ascribed phenomenological and neurobiological similarities to dream states in schizophrenia symptoms such as delusional beliefs, sensory hallucinations, instinctual behaviors, emotional disturbances, orientational instability, and bizarre imagery (Skrzypińska & Szmigielska, 2013; Hobson, Stickgold, & Pace-Schott, 1998). In both dream states and in acute schizophrenia, the person is involved in internal, cognitive events that are characterized by an incongruity and discontinuity of cognition and dream perception with rather limited connection to the outside world (Skrzypińska & Szmigielska, 2013; Hall, 1953; McCreery, 2008). Furthermore, a control mechanism that monitors the source of (internal or external) stimulation is lacking (Windt & Noreika, 2011). A study by Noreika, Valli, Markkula, Seppälä, and Revonsuo (2010) showed that dream states of patients with schizophrenia are even more bizarre compared to a non-clinical population. Looking at the neurobiological characteristics, there are striking similarities of schizophrenia and dream states (REM-sleep) from electrophysiological, topographic, and pharmacological approaches. In both cases, there is an impairment in inhibitory processes (Gottesmann, 2006), suppressed gamma rhythms in visual areas, prefrontal, and frontal cortices (Pérez-Garci, del Río-Portilla, Guevara, Arce, & Corsi-Cabrera, 2001). Similar alterations in cerebral blood flow and decreased activation of the dorsolateral prefrontal cortex might further explain disturbances in mentation and self-reflectiveness (Callicott et al., 2000; Maquet et al., 2000). The reduced thalamocortical gamma activity has previously been linked to the occurrence of hallucinations (Gottesmann, 2005; Behrendt & Young, 2004), higher dopamine activity during REM sleep might indicate an explanation for the loss of reflectiveness (Gottesmann, 2006; Gottesmann, 2005), and an increased activity of the Amygdala through higher levels of glutamate that might lead to problems with the perception of emotions (Gottesmann, 2006). Lastly, the levels of Noradrenaline, Serotonin, and Achetylcholine are decreased significantly in REM sleep and schizophrenia, while specifically Achetylcholine might be associated with hallucinations (Llewellyn, 2009) (for an in depth review of the similarities, see Skrzypińska & Szmigielska, 2013). Despite the striking parallels, there are still some unanswered questions. For example, it is yet unclear why in dreams visual hallucinations are more vivid, while in schizophrenia auditory hallucinations are more dominant (Skrzypińska & Szmigielska, 2013).

Patients with schizophrenia have also a specific deficit of sleep spindles (Manoach & Stickgold, 2019), i.e., neural oscillatory activity occurring during a stage of non-REM sleep and presumably mediating long-term memory consolidation. This deficit correlates with impaired sleep-dependent memory consolidation. Consolidated memories are malleable and can be destabilized and reconsolidated (Sinclair & Barense, 2018). The rate of consolidation seems to be driven by prediction errors. A surprising experience, when incongruent with prior knowledge, destabilizes episodic memories and promotes its updating (Sinclair & Barense, 2018). Experiments have shown that prediction error-driven memory consolidation improves learning performance also in artificial systems (Schillaci, Schmidt, & Miranda, 2020). Evidence links sleep-dependent memory consolidation with dreaming (Wamsley, 2014), as well as suggests that novel experience influences dream content especially in the visual domain (de Koninck, Christ, Hébert, & Rinfret, 1990; Kussé, Shaffii-LE Bourdiec, Schrouff, Matarazzo, & Maquet, 2012; Wamsley, Tucker, Payne, Benavides, & Stickgold, 2010). One claim is that delusions and dreams can both be seen as states with a deficient "reality testing" (Gerrans, 2014). Specifically, the model proposes that hallucinations and false memories emerge from faulty reality testing (Moulin, 2013; Bentall, 2003; Hobson, 1999), while dreams can be seen as hyperassociative default processing resulting from activation of the default mode network (DMN) that instantiates "raw material" for delusions, confabulations, and narrative context for waking cognition. In a healthy brain, these confabulations can be overwritten and (dis-) confirmed by evidence, which is thought to be associated with activity in the right dorsolateral prefrontal cortex (Gerrans, 2014). However, if these circuits are lesioned or hypoactive, as it may be in patients suffering from schizophrenia, the hypothesis cannot be overwritten leading to a missing evaluation of beliefs (Gerrans, 2014). This link between dreams and delusions explains the contextualization of beliefs which could explain the emergence of delusions. Since these confabulations cannot be contextualized as confabulations or hypotheses, and implausible beliefs cannot be overwritten, this might explain why patients with schizophrenia experience psychotic symptoms due to decontextualized memories (Gerrans, 2014).

In AI research, deep generative models have already been of use to simulate dream states (e.g. Google deepdream). Such neural networks were trained to detect patterns in pictures even when there is a high variance involved. By integrating a variational architecture with the iterative enhancement of the activation of certain layers of the network, this leads to the creation of pictures that look alien and far from reality the more iterative cycles were involved. Over-excitation of neurons in similar generative models have shown to lead to artificial hallucinations (Reichert, Series, & Storkey, 2013). Some of these "dreaming AIs" are based on aberrant salience attribution which is similar to psychotic states as explained above. Therefore, deep dreaming neural networks have been proposed as mechanisms representing the pathogenesis of schizophrenia that can be used to generate and test predictions for psychosis (Keshavan & Sudarshan, 2017). Limited research has focused on implementing generative models in robots for studying similar phenomena. As multi-modal embodied agents, robots are perfect test-beds for these investigations. In fact, dreaming is not a uni-modal phenomenon, as it is often associated with strong sensorimotor activity (Hobson, Pace-Schott, & Stickgold, 2000; Speth & Speth, 2018). Empirical studies have shown that motor imagery can be even induced during REM sleep through transcranial direct current stimulation (tDCS) in the motor cortex (Speth & Speth, 2016). Studying multi-modal generative models and memory systems in robots could provide insights on the nature of dreams and hallucinations.

6. Conclusions

In this review, we aimed to unravel the interactions between different informational sources in the construction of the self, and how to synthesize them into robotic agents. By modeling self-disorders, one can better understand disorders of self in a computational sense (computational psychiatry). For that, we discussed several models (e.g. circular inference, predictive coding). These allow us to

furnish a framework for possible explanations as to how and why some humans feel a disruption in their stream of consciousness or a demarcation from their body. We reported evidence for different models from computational and empirical human studies, and assessed their biological plausibility.

Some evidence points to the notion that neuronal Bayesian dynamics might encode expected change in the environment (Hohwy & Seth, 2020; Hohwy, Paton, & Palmer, 2016). We stress our core belief that the main computational foundations can be applied to both biological and artificial agents, and lead to the reduction of uncertainty and top-down messaging by a probabilistic, generative model (Hohwy & Seth, 2020). Utilizing predictive coding in human studies seems promising for capturing a fuller phenomenological picture of the self, by matching neural correlates and neural computations with the underlying conscious experience (Hohwy & Seth, 2020). In addition, employing this framework in cognitive developmental robotics shows promise for developing artificial agents with more advanced self-models (Georgie et al., 2019). These artificial self-models can be utilized for human phenomenological research of the self, or disturbances of the self. Specifically, predictive coding can explain multimodal integration through Bayesian optimal approximation and offers explanations for altered sense of ownership and sense of agency, as well as for hallucinations and delusions in clinical populations (e.g. overusing perceptual priors).

One strength of the predictive coding framework is that the phenomenology of being an embodied agent with a self concept is arguably created through active inference of an active self and its interoceptive states (Gallagher, 2000; Northoff, 2013; Neisser, 1988; Jeannerod, 2007; Jardri & Deneve, 2013; Huq et al., 1988; Lanillos, Dean-Leon, & Cheng, 2017; Dayan, Hinton, Neal, & Zemel, 1995; Tsakiris, 2017; Hafner et al., 2020). Also, inferences might be prompted by a regulation of hidden causes of states, rather than by an accurate representation (Hohwy & Seth, 2020; Seth, 2015a; Seth, 2015b; Wiese, 2014). These mechanisms, currently mostly applied to the study of biological agents, can be simulated and analyzed in cognitive developmental robotics with the added value of having access to the "black box". Understanding neurocognitive processes on phenomenology and investigating their impairments in psychiatric disorders might bring about newer approaches regarding the emergence of the self in humans, and spark ideas on how to develop a more sophisticated self model in embodied artificial agents.

Author Contributions

T.J.M.: Conceptualization, Writing—original draft; Y.K.G.: Conceptualization, Writing—review & editing; G.S.: Conceptualization, Writing—review & editing; M.V. and V.V.H. and L.K.: Conceptualization, Writing—review & editing, Supervision; All authors have read and agreed to the published version of the manuscript.

Funding

The work of T.J.M., L.K., and M.V. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the project "Functional aspects of the minimal self—the case of schizophrenia" (DFG KA 4920/1–1 VO 1744/2–1). The work of Y.K.G. and V.V.H. was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), in the project "Prerequisites for the Development of an Artificial Self" (402790442). Both projects are within the Special Priority Program "SPP—The Active Self" (SPP 2134). G.S. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 838861 ("Predictive Robots"). Predictive Robots is an associated project of the "SPP—The Active Self" (SPP 2134).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We want to thank both anonymous reviewers whose comments and critique helped to further improve the quality of this manuscript

References

Adams, R. A., Brown, H. R., & Friston, K. J. (2014). Bayesian inference, predictive coding and delusions. AVANT.J. *Philos. Int. Vanguard, 5*, 51–88.
Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry, 87*, 53–63.
Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry, 4*, 47.
Addington, J., Addington, D., & Maticka-Tyndale, E. (1991). Cognitive functioning and positive and negative symptoms in schizophrenia. *Schizophrenia research, 5*, 123–134.
Adery, L. H., Ichinose, M., Torregrossa, L. J., Wade, J., Nichols, H., Bekele, E., Bian, D., Gizdic, A., Granholm, E., Sarkar, N., et al. (2018). The acceptability and feasibility of a novel virtual reality based social skills training game for schizophrenia: Preliminary findings. *Psychiatry research, 270*, 496–502.
Aitchison, L., & Lengyel, M. (2017). With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology, 46*, 219–227.
Akins, K. A., & Dennett, D. C. (1986). Who may i say is calling? *Behavioral and Brain Sciences, 9*, 517–518.
Alderson-Day, B., Lima, C. F., Evans, S., Krishnan, S., Shanmugalingam, P., Fernyhough, C., & Scott, S. K. (2017). Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. *Brain, 140*, 2475–2489.
Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese, 195*, 2459–2482.

Amad, A., Cachia, A., Gorwood, P., Pins, D., Delmaire, C., Rolland, B., Mondino, M., Thomas, P., & Jardri, R. (2014). The multimodal connectivity of the hippocampal complex in auditory and visual hallucinations. *Molecular psychiatry, 19*, 184–191.
Andreasen, N. C., Arndt, S., Alliger, R., Miller, D., & Flaum, M. (1995). Symptoms of schizophrenia: Methods, meanings, and mechanisms. *Archives of general psychiatry, 52*, 341–351.
Aru, J., Rutiku, R., Wibral, M., Singer, W., & Melloni, L. (2016). Early effects of previous experience on conscious perception. Neuroscience of consciousness, 2016, niw004.
Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., & Yoshida, C. (2009). Cognitive developmental robotics: A survey. *IEEE transactions on autonomous mental development, 1*, 12–34.
Asada, M., MacDorman, K. F., Ishiguro, H., & Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous systems, 37*, 185–193.
Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain research reviews, 22*, 229–244.
Barber, M. J., Clark, J. W., & Anderson, C. H. (2003). Neural representation of probabilistic information. *Neural Computation, 15*, 1843–1864.
Barker, A. T., Jalinous, R., & Freeston, I. L. (1985). Non-invasive magnetic stimulation of human motor cortex. *The Lancet, 325*, 1106–1107.
Baron, R., Binder, A., & Wasner, G. (2010). Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. *The Lancet Neurology, 9*, 807–819.
Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*, 695–711.
Baumeister, R. F., & Exline, J. J. (2002). Mystical self loss: A challenge for psychological theory. *The International Journal for the Psychology of Religion, 12*, 15–20.
Bayes, Thomas (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London, 53*, 370–418. https://doi.org/10.1098/rstl.1763.0053.
Bays, P. M., & Wolpert, D. M. (2007). Computational principles of sensorimotor control that minimize uncertainty and variability. *The Journal of physiology, 578*, 387–396.
Bays, P. M., Wolpert, D. M., & Flanagan, J. R. (2005). Perception of the consequences of self-action is temporally tuned and event driven. *Current Biology, 15*, 1125–1128.
Bays, P. M., Wolpert, D. M., Haggard, E. P., Rosetti, Y., & Kawato, M. (2008). Predictive attenuation in the perception of touch. *Sensorimotor foundations of higher cognition, 22*, 339–358.
Bechtle, S., Schillaci, G., & Hafner, V. V. (2016). On the sense of agency and of object permanence in robots. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 166–171). IEEE.
Behrendt, R.-P., & Young, C. (2004). Hallucinations in schizophrenia, sensory impairment, and brain disease: A unifying model. *Behavioral and Brain sciences, 27*, 771.
Bell, J. S. (1964). On the einstein podolsky rosen paradox. *Physics Physique Fizika, 1*, 195.
Benrimoh, D., Parr, T., Vincent, P., Adams, R. A., & Friston, K. (2018). Active inference and auditory hallucinations. *Computational Psychiatry, 2*, 183–204.
Bentall, R. P. (2003). *Madness explained: Psychosis and human nature*. Penguin UK.
Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of cognitive neuroscience, 11*, 551–559.
Blakemore, S.-J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself? *Neuroreport, 11*, R11–R16.
Blanke, O., Landis, T., Spinelli, L., & Seeck, M. (2004). Out-of-body experience and autoscopy of neurological origin. *Brain, 127*, 243–258.
Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in cognitive sciences, 13*, 7–13.
Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., Landis, T., & Thut, G. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *Journal of Neuroscience, 25*, 550–557.
Blanke, O., Ortigue, S., Landis, T., & Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature, 419*, 269–270.
Boly, M., Moran, R., Murphy, M., Boveroux, P., Bruno, M.-A., Noirhomme, Q., Ledoux, D., Bonhomme, V., Brichant, J.-F., Tononi, G., et al. (2012). Connectivity changes underlying spectral eeg changes during propofol-induced loss of consciousness. *Journal of Neuroscience, 32*, 7082–7090.
Bradley, F. H. (2016). *Appearance and reality: a metaphysical essay*. Routledge.
Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H., & Philipsen, A. (2018). The senses of agency and ownership: a review. *Frontiers in psychology, 9*, 535.
Brewin, C. R., Gregory, J. D., Lipton, M., & Burgess, N. (2010). Intrusive images in psychological disorders: characteristics, neural mechanisms, and treatment implications. *Psychological review, 117*, 210.
Callicott, J. H., Bertolino, A., Mattay, V. S., Langheim, F. J., Duyn, J., Coppola, R., Goldberg, T. E., & Weinberger, D. R. (2000). Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited. *Cerebral cortex, 10*, 1078–1092.
Campbell, J. (1999). Schizophrenia, the space of reasons, and thinking as a motor process. *The Monist, 82*, 609–625.
Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots*. MIT press.
Cho, R., Konecky, R., & Carter, C. S. (2006). Impairments in frontal cortical γsynchrony and cognitive control in schizophrenia. *Proceedings of the National Academy of Sciences, 103*, 19878–19883.
Christoff, K., Cosmelli, D., Legrand, D., & Thompson, E. (2011). Specifying the self for cognitive neuroscience. *Trends in cognitive sciences, 15*, 104–112.
Ciria, A., Schillaci, G., Pezzulo, G., Hafner, V.V., & Lara, B. (2021). Predictive processing in cognitive robotics: a review. arXiv:(to appear in) MIT Press journal on Neural Computation. arXiv 2101.06611.
Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences, 36*, 181–204.
Corlett, P. R., Honey, G. D., Krystal, J. H., & Fletcher, P. C. (2011). Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology, 36*, 294–315.
Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R., III (2019). Hallucinations and strong priors. *Trends in cognitive sciences, 23*, 114–127.
Corlett, P. R., Krystal, J. H., Taylor, J. R., & Fletcher, P. C. (2009). Why do delusions persist? *Frontiers in human neuroscience, 3*, 12.
Corlett, P. R., Taylor, J., Wang, X.-J., Fletcher, P., & Krystal, J. (2010). Toward a neurobiology of delusions. *Progress in neurobiology, 92*, 345–369.
Craig, A. D. (2002). How do you feel? interoception: the sense of the physiological condition of the body. *Nature reviews neuroscience, 3*, 655–666.
Craig, A. D., & Craig, A. (2009). How do you feel–now? the anterior insula and human awareness. *Nature reviews neuroscience, 10*.
Damasio, A. (2003). Mental self: The person within. *Nature, 423*, 227.
Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.
David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and cognition, 17*, 523–534.
Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation, 7*, 889–904.
De Haan, S., & Fuchs, T. (2010). The ghost in the machine: disembodiment in schizophrenia–two case studies. *Psychopathology, 43*, 327–333.
De Ridder, D., Vanneste, S., & Freeman, W. (2014). The bayesian brain: phantom percepts resolve sensory uncertainty. *Neuroscience & Biobehavioral Reviews, 44*, 4–15.
Deane, G., Miller, M. D., & Wilkinson, S. (2020). Losing ourselves: Active inference, depersonalization and meditation. *Frontiers in Psychology, 11*, 2893.
Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: a social cognitive neuroscience view. *Trends in cognitive sciences, 7*, 527–533.
Denève, S., & Jardri, R. (2016). Circular inference: mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences, 11*, 40–48.
d'Espagnat, B. (1979). The quantum theory and reality. *Scientific American, 241*, 158–181.
Di Paolo, E. A., & Thompson, E. (2014). *The enactive approach. The Routledge handbook of embodied cognition* (pp. 68–78).
Dogge, M., Custers, R., & Aarts, H. (2019). Moving forward: On the limits of motor-based forward models. *Trends in Cognitive Sciences, 23*, 743–753.
Ebisch, S. J., & Gallese, V. (2015). A neuroscientific perspective on the nature of altered self-other relationships in schizophrenia. *Journal of Consciousness Studies, 22*, 220–240.
Erikson, E.H. (1950). Childhood and society, new york (ww norton) 1950.

Farrer, C., Franck, N., Frith, C. D., Decety, J., Georgieff, N., d'Amato, T., & Jeannerod, M. (2004). Neural correlates of action attribution in schizophrenia. *Psychiatry Research: Neuroimaging, 131*, 31–44.
Fee, M. S. (2014). The role of efference copy in striatal learning. *Current opinion in neurobiology, 25*, 194–200.
Feinberg, I. (1978). Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophrenia bulletin, 4*, 636.
Feinberg, T.E. (2011). Brain and self: bridging the gap. Consciousness and cognition (Print), 20.
Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10*, 48–58.
Flor, H., Nikolajsen, L., & Jensen, T. S. (2006). Phantom limb pain: a case of maladaptive cns plasticity? *Nature reviews neuroscience, 7*, 873–881.
Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16*, 1325–1352.
Friston, K. (2005). A theory of cortical responses. *Phil. trans. of the Royal Society B: Biological sciences, 360*, 815–836.
Friston, K. (2012). The history of the future of the bayesian brain. *NeuroImage, 62*, 1230–1233.
Friston, K. (2012). Prediction, perception and agency. *International Journal of Psychophysiology, 83*, 248–252.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural computation, 29*, 1–49.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 1211–1221.
Friston, K., Samothrakis, S., & Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biological cybernetics, 106*, 523–541.
Friston, K. J., & Frith, C. D. (1995). Schizophrenia: a disconnection syndrome. *Clin Neurosci, 3*, 89–97.
Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry, 1*, 148–158.
Frith, C. (2012). Explaining delusions of control: The comparator model 20 years on. *Consciousness and cognition, 21*, 52–54.
Frith, C. D. (1979). Consciousness, information processing and schizophrenia. *British Journal of.*
Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. Philosophical Transactions of the Royal Society of London. *Series B: Biological Sciences, 355*, 1771–1788.
Fuchs, T. (2005). Corporealized and disembodied minds: a phenomenological view of the body in melancholia and schizophrenia. *Philosophy, Psychiatry, & Psychology, 12*, 95–107.
Fuchs, T. (2009). Embodied cognitive neuroscience and its consequences for psychiatry. *Poiesis & Praxis, 6*, 219–233.
Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences, 4*, 14–21.
Gallagher, S. (2004). Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology, 37*, 8–19.
Gallagher, S. (2006). *How the body shapes the mind*. Clarendon Press.
Gallagher, S., & Zahavi, D. (2016). Phenomenological approaches to self-consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
Gallese, V., & Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia, 48*, 746–755.
Gama, F., Shcherban, M., Rolf, M., & Hoffmann, M. (2020). Active exploration for body model learning through self-touch on a humanoid robot with artificial skin. arXiv:2008.13483.
Garety, P. A., & Hemsley, D. R. (1997). *Delusions: Investigations into the psychology of delusional reasoning* (volume 36). Psychology Press.
Georgie, Y. K., Schillaci, G., & Hafner, V. V. (2019). An interdisciplinary overview of developmental indices and behavioral measures of the minimal self. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 129–136). IEEE.
Gerrans, P. (2014). Pathologies of hyperfamiliarity in dreams, delusions and déjà vu. *Frontiers in Psychology, 5*, 97.
Gerrans, P. (2019). Depersonalization disorder, affective processing and predictive coding. *Review of Philosophy and Psychology, 10*, 401–418.
Giersch, A., & Mishara, A. L. (2017). Is schizophrenia a disorder of consciousness? experimental and phenomenological support for anomalous unconscious processing. *Frontiers in psychology, 8*, 1659.
Gold, K., & Scassellati, B. (2009). Using probabilistic reasoning over time to self-recognize. *Robotics and autonomous systems, 57*, 384–392.
Gordon, N., Tsuchiya, N., Koenig-Robert, R., & Hohwy, J. (2019). Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. *PLoS biology, 17*, e3000233.
Gottesmann, C. (2005). Dreaming and schizophrenia: a common neurobiological background. *Sleep and biological rhythms, 3*, 64–74.
Gottesmann, C. (2006). The dreaming sleep stage: a new neurobiological model of schizophrenia? *Neuroscience, 140*, 1105–1115.
Gray, J. A., Feldon, J., Rawlins, J., Hemsley, D., & Smith, A. (1991). The neuropsychology of schizophrenia. *Behavioral and Brain Sciences, 14*, 1–20.
Griffiths, O., Langdon, R., Le Pelley, M. E., & Coltheart, M. (2014). Delusions and prediction error: re-examining the behavioural evidence for disrupted error signalling in delusion formation. *Cognitive neuropsychiatry, 19*, 439–467.
Hafner, V. V., Loviken, P., Pico Villalpando, A., & Schillaci, G. (2020). Prerequisites for an artificial self. *Frontiers in Neurorobotics, 14*.
Hall, C. S. (1953). A cognitive theory of dreams. *The Journal of General Psychology, 49*, 273–282.
Heider, B. (2000). Visual form agnosia: neural mechanisms and anatomical foundations. *Neurocase, 6*, 1–12.
Heinz, A. (2002). Dopaminergic dysfunction in alcoholism and schizophrenia–psychopathological and behavioral correlates. *European Psychiatry, 17*, 9–16.
Heinz, A. (2014). *Der Begriff der psychischen Krankheit*. Suhrkamp Verlag.
Heinz, A., Murray, G. K., Schlagenhauf, F., Sterzer, P., Grace, A. A., & Waltz, J. A. (2019). Towards a unifying cognitive, neurophysiological, and computational neuroscience account of schizophrenia. *Schizophrenia Bulletin, 45*, 1092–1100.
Helmholtz, Hermann (1867), *9. Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*. Leipzig, Germany: Voss.
Hermle, L., Fünfgeld, M., Oepen, G., Botsch, H., Borchardt, D., Gouzoulis, E., Fehrenbach, R. A., & Spitzer, M. (1992). Mescaline-induced psychopathological, neuropsychological, and neurometabolic effects in normal subjects: experimental psychosis as a tool for psychiatric research. *Biological psychiatry, 32*, 976–991.
Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation, 33*, 1–49.
Hinz, N.-A., Lanillos, P., Mueller, H., & Cheng, G. (2018). Drifting perceptual patterns suggest prediction errors fusion rather than hypothesis selection: replicating the rubber-hand illusion on a robot. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 125–132). IEEE.
Hobson, J. A. (1999). *Dreaming as delirium: How the brain goes out of its mind*. MIT Press.
Hobson, J. A., Pace-Schott, E. F., & Stickgold, R. (2000). Dreaming and the brain: toward a cognitive neuroscience of conscious states. *Behavioral and brain sciences, 23*, 793–842.
Hobson, J. A., Stickgold, R., & Pace-Schott, E. F. (1998). The neuropsychology of rem sleep dreaming. *Neuroreport, 9*, R1–R14.
Hohwy, J., Paton, B., & Palmer, C. (2016). Distrusting the present. *Phenomenology and the Cognitive Sciences, 15*, 315–335.
Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness.
Hommel, B. (2013). Ideomotor action control: On the perceptual grounding of voluntary actions and agents. Action science: Foundations of an emerging discipline, (pp. 113–136).
Horan, W. P., Green, M. F., Knowlton, B. J., Wynn, J. K., Mintz, J., & Nuechterlein, K. H. (2008). Impaired implicit learning in schizophrenia. *Neuropsychology, 22*, 606.
Horga, G., & Abi-Dargham, A. (2019). An integrative framework for perceptual disturbances in psychosis. *Nature Reviews Neuroscience*, 1–16.
Howes, O., Bose, S., Turkheimer, F., Valli, I., Egerton, A., Stahl, D., Valmaggia, L., Allen, P., Murray, R., & McGuire, P. (2011). Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a pet study. *Molecular psychiatry, 16*, 885–886.
Humpston, C. S., & Broome, M. R. (2016). The spectra of soundless voices and audible thoughts: Towards an integrative model of auditory verbal hallucinations and thought insertion. *Review of Philosophy and Psychology, 7*, 611–629.
Humpston, C.S., & Broome, M.R. (2020). Thinking, believing, and hallucinating self in schizophrenia. The Lancet Psychiatry.
Huq, S., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A, 40*, 801–812.

Insel, T. R. (2010). Rethinking schizophrenia. *Nature, 468*, 187–193.
Insel, T.R., & Lieberman, J.A. (2013). Dsm-5 and rdoc: shared interests.
Iwabuchi, S. J., & Palaniyappan, L. (2017). Abnormalities in the effective connectivity of visuothalamic circuitry in schizophrenia. *Psychological medicine, 47*, 1300.
Jardri, R., & Deneve, S. (2013). Circular inferences in schizophrenia. *Brain, 136*, 3227–3241.
Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., Smailes, D., Sterzer, P., Corlett, P. R., Leptourgos, P., et al. (2016). Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? *Schizophrenia bulletin, 42*, 1124–1134.
Jeannerod, M. (2007). Being oneself. *Journal of Physiology-Paris, 101*, 161–168.
Jeannerod, M. (2009). The sense of agency and its disturbances in schizophrenia: a reappraisal. *Experimental Brain Research, 192*, 527.
Jeannerod, M., & Johnson-Frey, S. (2003). Simulation of action as a unifying concept for motor cognition. Taking action: Cognitive neuroscience perspectives on intentional acts, (pp. 139–163).
Jones, C., Watson, D., & Fone, K. (2011). Animal models of schizophrenia. *British journal of pharmacology, 164*, 1162–1194.
Kaminski, J., Sterzer, P., & Mishara, A. (2019). "seeing rain": integrating phenomenological and bayesian predictive coding approaches to visual hallucinations and self-disturbances (ichstörungen) in schizophrenia. *Consciousness and cognition, 73*, 102757.
Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *American journal of Psychiatry, 160*, 13–23.
Keshavan, M. S., & Sudarshan, M. (2017). Deep dreaming, aberrant salience and psychosis: connecting the dots by artificial neural networks. *Schizophrenia research, 188*, 178–181.
Khrennikov, A. (2004). Probabilistic pathway representation of cognitive information. *Journal of theoretical biology, 231*, 597–613.
Kilteni, K., & Ehrsson, H. H. (2017). Body ownership determines the attenuation of self-generated tactile sensations. *Proceedings of the National Academy of Sciences, 114*, 8426–8431.
Kilteni, K., Maselli, A., Kording, K. P., & Slater, M. (2015). Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in human neuroscience, 9*, 141.
Kiverstein, J. (2020). Free energy and the self: an ecological–enactive interpretation. *Topoi, 39*, 559–574.
Kneissler, J., Drugowitsch, J., Friston, K., & Butz, M. V. (2015). Simultaneous learning and filtering without delusions: A bayes-optimal derivation of combining predictive inference and adaptive filtering. *Frontiers in computational neuroscience, 9*, 47.
Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences, 27*, 712–719.
Knyazev, G. (2013). Eeg correlates of self-referential processing. *Frontiers in human neuroscience, 7*, 264.
de Koninck, J., Christ, G., Hébert, G., & Rinfret, N. (1990). Language learning efficiency, dreams and rem sleep. *Psychiatric Journal of the University of Ottawa*.
Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry, 3*, 935–946.
Kussé, C., Shaffii-LE Bourdiec, A., Schrouff, J., Matarazzo, L., & Maquet, P. (2012). Experience-dependent induction of hypnagogic images during daytime naps: A combined behavioural and eeg study. *Journal of Sleep Research, 21*, 10–20.
Lang, C., Schillaci, G., & Hafner, V. V. (2018). A deep convolutional neural network model for sense of agency and object permanence in robots. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 257–262). IEEE.
Lanillos, P., & Cheng, G. (2018). Adaptive robot body learning and estimation through predictive coding. arXiv preprint arXiv:1805.03104.
Lanillos, P., Dean-Leon, E., & Cheng, G. (2016). Yielding self-perception in robots through sensorimotor contingencies. *IEEE Transactions on Cognitive and Developmental Systems, 9*, 100–112.
Lanillos, P., Dean-Leon, E., & Cheng, G. (2017). Enactive self: a study of engineering perspectives to obtain the sensorimotor self through enaction. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 72–78). IEEE.
Lanillos, P., Oliva, D., Philippsen, A., Yamashita, Y., Nagai, Y., & Cheng, G. (2020a). A review on neural network models of schizophrenia and autism spectrum disorder. *Neural Networks, 122*, 338–363.
Lanillos, P., Pagès, J., & Cheng, G. (2020b). Robot self/other distinction: active inference meets neural networks learning in a mirror. In ECAI.
Leptourgos, P., Denève, S., & Jardri, R. (2017). Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Current opinion in neurobiology, 46*, 154–161.
Liang, M., Zhou, Y., Jiang, T., Liu, Z., Tian, L., Liu, H., & Hao, Y. (2006). Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *Neuroreport, 17*, 209–213.
Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in human neuroscience, 7*, 547.
Lisman, J. E., & Grace, A. A. (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron, 46*, 703–713.
Llewellyn, S. (2009). In two minds? is schizophrenia a state 'trapped'between waking and dreaming? *Medical hypotheses, 73*, 572–579.
Lubow, R., & Moore, A. (1959). Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of comparative and physiological psychology, 52*, 415.
Lubow, R. E. (1973). Latent inhibition. *Psychological bulletin, 79*, 398.
Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection science, 15*, 151–190.
Ma, K., & Hommel, B. (2015a). Body-ownership for actively operated non-corporeal objects. *Consciousness and cognition, 36*, 75–86.
Ma, K., & Hommel, B. (2015b). The role of agency for perceived ownership in the virtual hand illusion. *Consciousness and cognition, 36*, 277–288.
Maffei, A. (2017). Fifty shades of inhibition. *Current opinion in neurobiology, 43*, 43–47.
Maher, B. (1983). A tentative theory of schizophrenic utterance. *Progress in experimental personality research, 12*, 1–52.
Manoach, D. S., & Stickgold, R. (2019). Abnormal sleep spindles, memory consolidation, and schizophrenia. *Annual review of clinical psychology, 15*, 451–479.
Maquet, P., et al. (2000). Functional neuroimaging of normal human sleep by positron emission tomography. *Journal of sleep research, 9*, 207–232.
Marcotte, E. R., Pearson, D. M., & Srivastava, L. K. (2001). Animal models of schizophrenia: a critical review. *Journal of psychiatry & neuroscience*.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information* (p. 2). New York, NY: henry holt and co. Inc..
Mason, N., Kuypers, K., Müller, F., Reckweg, J., Tse, D., Toennes, S., Hutten, N., Jansen, J., Stiers, P., Feilding, A., et al. (2020). Me, myself, bye: regional alterations in glutamate and the experience of ego dissolution with psilocybin. *Neuropsychopharmacology*, 1–11.
McCormick, D. A., Wang, Z., & Huguenard, J. (1993). Neurotransmitter control of neocortical neuronal activity and excitability. *Cerebral cortex, 3*, 387–398.
McCreery, C. (2008). *Dreams and psychosis: A new look at an old hypothesis*. Oxford Forum.
McCutcheon, R. A., Abi-Dargham, A., & Howes, O. D. (2019). Schizophrenia, dopamine and the striatum: from biology to symptoms. *Trends in neurosciences, 42*, 205–220.
van der Meer, L., Costafreda, S., Aleman, A., & David, A. S. (2010). Self-reflection and the brain: a theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. *Neuroscience & Biobehavioral Reviews, 34*, 935–946.
Metzinger, T. (2014). *Der Ego-Tunnel: Eine neue Philosophie des Selbst*. Von der Hirnforschung zur Bewusstseinsethik. Piper Verlag.
Michalareas, G., Vezoli, J., Van Pelt, S., Schoffelen, J.-M., Kennedy, H., & Fries, P. (2016). Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron, 89*, 384–397.
Michels, F., Schilling, C., Rausch, F., Eifler, S., Zink, M., Meyer-Lindenberg, A., & Schredl, M. (2014). Nightmare frequency in schizophrenic patients, healthy relatives of schizophrenic patients, patients at high risk states for psychosis, and healthy controls. *International Journal of Dream Research.*
Miller, R. (1976). Schizophrenic psychology, associative learning and the role of forebrain dopamine. *Medical Hypotheses, 2*, 203–211.
Møller, P., & Husby, R. (2000). The initial prodrome in schizophrenia: searching for naturalistic core dimensions of experience and behavior. *Schizophrenia Bulletin, 26*, 217–232.
Möller, T. J., Braun, N., Thöne, A.-K., Herrmann, C. S., & Philipsen, A. (2020). The senses of agency and ownership in patients with borderline personality disorder. *Frontiers in Psychiatry, 11*, 474.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences, 16*, 72–80.
Moulin, C. J. (2013). Disordered recognition memory: recollective confabulation. *Cortex, 49*, 1541–1552.
Mückenheim, W. (1983). Das epr-paradoxon und die unbestimmtheit der realität. *Physikalische Blätter, 39*, 331–336.
Murray, J. D., Anticevic, A., Gancsos, M., Ichinose, M., Corlett, P. R., Krystal, J. H., & Wang, X.-J. (2014). Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cerebral cortex, 24*, 859–872.
Murray, R. M., & Lewis, S. W. (1987). Is schizophrenia a neurodevelopmental disorder? *British medical journal (Clinical research ed.), 295*, 681.
National Collaborating Centre for Mental Health (UK and others) (2014). Psychosis and schizophrenia in adults: treatment and management.
Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical psychology, 1*, 35–59.
Nguyen, P.D., Georgie, Y.K., Kayhan, E., Eppe, M., Hafner, V.V., & Wermter, S. (2020). Sensorimotor representation learning for an active self in robots: A model survey. arXiv preprint arXiv:2011.12860.
Noreika, V., Valli, K., Markkula, J., Seppälä, K., & Revonsuo, A. (2010). Dream bizarreness and waking thought in schizophrenia. *Psychiatry Research, 178*, 562–564.
Northoff, G. (2013). Brain and self–a neurophilosophical account. *Child and adolescent psychiatry and mental health, 7*, 28.
Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain–a meta-analysis of imaging studies on the self. *Neuroimage, 31*, 440–457.
Notredame, C.-E., Pins, D., Deneve, S., & Jardri, R. (2014). What visual illusions teach us about schizophrenia. *Frontiers in integrative neuroscience, 8*, 63.
Owen, M. J., O'Donovan, M. C., Thapar, A., & Craddock, N. (2011). Neurodevelopmental hypothesis of schizophrenia. *The British journal of psychiatry, 198*, 173–175.
Parnas, J., & Handest, P. (2003). Phenomenology of anomalous self-experience in early schizophrenia. *Comprehensive psychiatry, 44*, 121–134.
Parnas, J., Møller, P., Kircher, T., Thalbitzer, J., Jansson, L., Handest, P., & Zahavi, D. (2005). Ease: examination of anomalous self-experience. *Psychopathology, 38*, 236.
Parnas, J., & Sass, L. A. (2001). Self, solipsism, and schizophrenic delusions. *Philosophy, Psychiatry, & Psychology, 8*, 101–120.
Parr, T., Rees, G., & Friston, K. J. (2018). Computational neuropsychology and bayesian inference. *Frontiers in human neuroscience, 12*, 61.
Pérez-Garci, E., del Río-Portilla, Y., Guevara, M. A., Arce, C., & Corsi-Cabrera, M. (2001). Paradoxical sleep is characterized by uncoupled gamma activity between frontal and perceptual cortical regions. *Sleep, 24*, 118–126.
Persinger, M. A., & Healey, F. (2002). Experimental facilitation of the sensed presence: Possible intercalation between the hemispheres induced by complex magnetic fields. *The Journal of nervous and mental disease, 190*, 533–541.
Petzschner, F. H., Weber, L. A., Gard, T., & Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: toward a joint framework for differential diagnosis. *Biological Psychiatry, 82*, 421–430.
Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT press.
Philippsen, A., & Nagai, Y. (2020a). Deficits in prediction ability trigger asymmetries in behavior and internal representation. *Frontiers in psychiatry, 11*, 1253.
Philippsen, A., & Nagai, Y. (2020b). A predictive coding account for cognition in human children and chimpanzees: a case study of drawing. In *IEEE Transactions on Cognitive and Developmental Systems*.
Piaget, J. (1954). The construction of reality in the child (m. cook, trans.). new york, ny, us.
Pickering, M. J., & Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. *Trends in cognitive sciences, 18*, 451–456.
Pico, A., Schillaci, G., Hafner, V. V., & Lara, B. (2016). How do i sound like? forward models for robot ego-noise prediction. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 246–251). IEEE.
Popper, K.R. (1959). The logic of scientific discovery hutchinson. Hughes, John, (1987)."La Filosofía de la Investigación Social", Breviarios, Fondo de Cultura Económica, México.
Powers, A. R., Mathys, C., & Corlett, P. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science, 357*, 596–600.
Powers, A. R., III, Kelley, M. S., & Corlett, P. R. (2017). Varieties of voice-hearing: psychics and the psychosis continuum. *Schizophrenia bulletin, 43*, 84–98.
Preller, K. H., Razi, A., Zeidman, P., Stämpfli, P., Friston, K. J., & Vollenweider, F. X. (2019). Effective connectivity changes in lsd-induced altered states of consciousness in humans. *Proceedings of the National Academy of Sciences, 116*, 2743–2748.
Prescott, T. J., & Camilleri, D. (2019). The synthetic psychology of the self. In *Cognitive architectures* (pp. 85–104). Springer.
Proust, J. (2006). Agency in schizophrenia from a control theory viewpoint.
Pynn, L. K., & DeSouza, J. F. (2013). The function of efference copy signals: implications for symptoms of schizophrenia. *Vision research, 76*, 124–133.
Ramachandran, V. S., Brang, D., & McGeoch, P. D. (2010). Dynamic reorganization of referred sensations by movements of phantom limbs. *Neuroreport, 21*, 727–730.
Ramachandran, V. S., & Rogers-Ramachandran, D. (1996). Synaesthesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 263*, 377–386.
Rao, R. P., Olshausen, B. A., & Lewicki, M. S. (2002). *Probabilistic models of the brain: Perception and neural function*. MIT press.
Rapoport, J., Giedd, J., & Gogtay, N. (2012). Neurodevelopmental model of schizophrenia: update 2012. *Molecular psychiatry, 17*, 1228–1238.
Rector, N. A., Beck, A. T., & Stolar, N. (2005). The negative symptoms of schizophrenia: a cognitive perspective. *The Canadian Journal of Psychiatry, 50*, 247–257.
Reichert, D. P., Series, P., & Storkey, A. J. (2013). Charles bonnet syndrome: evidence for a generative model in the cortex? *PLoS Comput Biol, 9*, e1003134.
Richer, F., Martinez, M., Robert, M., Bouvier, G., & Saint-Hilaire, J.-M. (1993). Stimulation of human somatosensory cortex: tactile and body displacement perceptions in medial regions. *Experimental brain research, 93*, 173–176.
Riva, G. (2018). The neuroscience of body memory: From the self through the space to the others. *Cortex, 104*, 241–260.
Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and cognition, 12*, 717–731.
Rolland, B., Amad, A., Poulet, E., Bordet, R., Vignaud, A., Bation, R., Delmaire, C., Thomas, P., Cottencin, O., & Jardri, R. (2015). Resting-state functional connectivity of the nucleus accumbens in auditory and visual hallucinations in schizophrenia. *Schizophrenia bulletin, 41*, 291–299.
Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a pet investigation of agency. *Nature neuroscience, 4*, 546–550.
Samad, M., Chung, A. J., & Shams, L. (2015). Perception of body ownership is driven by bayesian sensory inference. *PloS one, 10*, e0117178.
Sass, L. (2004). Affectivity in schizophrenia a phenomenological view. *Journal of consciousness studies, 11*, 127–147.
Sass, L. A., & Parnas, J. (2003). Schizophrenia, consciousness, and the self. *Schizophrenia bulletin, 29*, 427–444.
Sass, L. A., & Parnas, J. (2007). Explaining schizophrenia: the relevance of phenomenology. *Reconceiving schizophrenia*, 63–95.
Schillaci, G., Ritter, C.-N., Hafner, V.V., & Lara, B. (2016). Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents. In Proceedings of the Artificial Life Conference 2016 13 (pp. 390–397). MIT Press.
Schillaci, G., Schmidt, U., & Miranda, L. (2020). Prediction error-driven memory consolidation for continual learning. on the case of adaptive greenhouse models. arXiv preprint arXiv:2006.12616.
Schmack, K., de Castro, A. G.-C., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., Heinz, A., Petrovic, P., & Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *Journal of Neuroscience, 33*, 13701–13712.
Schmack, K., Rothkirch, M., Priller, J., & Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Human brain mapping, 38*, 1767–1779.
Schuman, C.D., Potok, T.E., Patton, R.M., Birdwell, J.D., Dean, M.E., Rose, G.S., & Plank, J.S. (2017). A survey of neuromorphic computing and neural networks in hardware. arXiv preprint arXiv:1705.06963.
Seth, A. (2015a). The cybernetic bayesian brain–from interoceptive inference to sensorimotor contingencies,(w:) open mind, red. t. metzinger, jm windt.
Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences, 17*, 565–573.
Seth, A. K. (2015b). *Inference to the best prediction. Open MIND*. Frankfurt am Main: MIND Group.
Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology, 2*, 395.
Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of science, 77*, 477–500.
Shergill, S. S., Bays, P. M., Frith, C. D., Wolpert, D. M., et al. (2003). Two eyes for an eye: the neuroscience of force escalation. *Science, 301*, 187.
Shergill, S. S., Samson, G., Bays, P. M., Frith, C. D., & Wolpert, D. M. (2005). Evidence for sensory prediction deficits in schizophrenia. *American Journal of Psychiatry, 162*, 2384–2386.

Shipp, S. (2017). The functional logic of corticostriatal connections. *Brain Structure and Function, 222*, 669–706.
Simpson, E. H., Kellendonk, C., & Kandel, E. (2010). A possible role for the striatum in the pathogenesis of the cognitive symptoms of schizophrenia. *Neuron, 65*, 585–596.
Sinclair, A. H., & Barense, M. D. (2018). Surprise and destabilize: prediction error influences episodic memory reconsolidation. *Learning & Memory, 25*, 369–381.
Skrzypińska, D., & Szmigielska, B. (2013). What links schizophrenia and dreaming? common phenomenological and neurobiological features of schizophrenia and rem sleep. *Archives of Psychiatry and Psychotherapy, 15*.
Speth, C., & Speth, J. (2018). A new measure of hallucinatory states and a discussion of rem sleep dreaming as a virtual laboratory for the rehearsal of embodied cognition. *Cognitive science, 42*, 311–333.
Speth, J., & Speth, C. (2016). Motor imagery in rem sleep is increased by transcranial direct current stimulation of the left motor cortex (c3). *Neuropsychologia, 86*, 57–65.
Stanghellini, G. (2004). *Disembodied spirits and deanimated bodies: The psychopathology of common sense*. Oxford University Press.
Stephan, K. E., Baldeweg, T., & Friston, K. J. (2006). Synaptic plasticity and dysconnection in schizophrenia. *Biological psychiatry, 59*, 929–939.
Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current opinion in neurobiology, 25*, 85–92.
Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. The MIT press.
Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological psychiatry, 84*, 634–643.
Sterzer, P., Mishara, A. L., Voss, M., & Heinz, A. (2016). Thought insertion as a self-disturbance: an integration of predictive coding and phenomenological approaches. *Frontiers in human neuroscience, 10*, 502.
Sterzer, P., Voss, M., Schlagenhauf, F., & Heinz, A. (2019). Decision-making in schizophrenia: a predictive-coding perspective. *Neuroimage, 190*, 133–143.
Stoytchev, A. (2009). Some basic principles of developmental robotics. *IEEE Transactions on Autonomous Mental Development, 1*, 122–130.
Stuke, H., Weilnhammer, V. A., Sterzer, P., & Schmack, K. (2019). Delusion proneness is linked to a reduced usage of prior beliefs in perceptual decisions. *Schizophrenia bulletin, 45*, 80–86.
Subedi, B., & Grossberg, G.T. (2011). Phantom limb pain: mechanisms and treatment approaches. Pain research and treatment, 2011.
Swerdlow, N. R., Braff, D. L., Hartston, H., Perry, W., & Geyer, M. A. (1996). Latent inhibition in schizophrenia. *Schizophrenia research, 20*, 91–103.
Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and cognition, 17*, 219–239.
Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: an interplay between prediction and postdiction. *Frontiers in psychology, 4*, 127.
Tandon, R., Nasrallah, H. A., & Keshavan, M. S. (2009). Schizophrenia,"just the facts" 4. clinical features and conceptualization. *Schizophrenia research, 110*, 1–23.
Tani, J., & White, J. (2020). Cognitive neurorobotics and self in the shared world, a focused review of ongoing research. *Adaptive Behavior*. p. 1059712320962158.
Tarantola, A. (2006). Popper, bayes and the inverse problem. *Nature physics, 2*, 492–494.
Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P. R., Goodyer, I. M., & Fletcher, P. C. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences, 112*, 13401–13406.
Thompson, E., & Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi, 28*, 23–30.
Timmermann, C., Roseman, L., Williams, L., Erritzoe, D., Martial, C., Cassol, H., Laureys, S., Nutt, D., & Carhart-Harris, R. (2018). Dmt models the near-death experience. *Frontiers in psychology, 9*, 1424.
Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*, 450–461.
Tsakiris, M. (2010). My body in the brain: a neurocognitive model of body-ownership. *Neuropsychologia, 48*, 703–712.
Tsakiris, M. (2017). The multisensory basis of the self: from body to identity to others. *The Quarterly Journal of Experimental Psychology, 70*, 597–609.
Tsakiris, M., Hesse, M. D., Boy, C., Haggard, P., & Fink, G. R. (2007). Neural signatures of body ownership: a sensory network for bodily self-consciousness. *Cerebral cortex, 17*, 2235–2244.
Valton, V., Romaniuk, L., Steele, J. D., Lawrie, S., & Seriès, P. (2017). Comprehensive review: computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews, 83*, 631–646.
Van Den Bos, E., & Jeannerod, M. (2002). Sense of body and sense of action both contribute to self-recognition. *Cognition, 85*, 177–187.
Van Den Buuse, M., Garner, B., Gogos, A., & Kusljic, S. (2005). Importance of animal models in schizophrenia research. *Australian & New Zealand Journal of Psychiatry, 39*, 550–557.
Varela, F. J., Thompson, E., & Rosch, E. (2016). *The embodied mind: Cognitive science and human experience*.
Verschoor, S. A., & Hommel, B. (2017). Self-by-doing: The role of action for self-acquisition. *Social Cognition, 35*, 127–145.
Vlisides, P., Bel-Bahar, T., Nelson, A., Chilton, K., Smith, E., Janke, E., Tarnal, V., Picton, P., Harris, R., & Mashour, G. (2018). Subanaesthetic ketamine and altered states of consciousness in humans. *British journal of anaesthesia, 121*, 249–259.
Vogeley, K., & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in cognitive sciences, 7*, 38–42.
Vosgerau, G., & Newen, A. (2007). Thoughts, motor actions, and the self. *Mind & Language, 22*, 22–43.
Voss, M., Chambon, V., Wenke, D., Kühn, S., & Haggard, P. (2017). In and out of control: brain mechanisms linking fluency of action selection to self-agency in patients with schizophrenia. *Brain, 140*, 2226–2239.
Voss, M., Ingram, J. N., Haggard, P., & Wolpert, D. M. (2006). Sensorimotor attenuation by central motor command signals in the absence of movement. *Nature neuroscience, 9*, 26–27.
Voss, M., Ingram, J. N., Wolpert, D. M., & Haggard, P. (2008). Mere expectation to move causes attenuation of sensory signals. *PLoS One, 3*, e2866.
Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., & Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain, 133*, 3104–3112.
Wamsley, E. J. (2014). Dreaming and offline memory consolidation. *Current Neurology and Neuroscience Reports, 14*, 433.
Wamsley, E. J., Tucker, M., Payne, J. D., Benavides, J. A., & Stickgold, R. (2010). Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. *Current Biology, 20*, 850–855.
Wegner, D. M. (2017). *The illusion of conscious will*. MIT Press.
Weinberger, D. R., Berman, K. F., & Zec, R. F. (1986). Physiologic dysfunction of dorsolateral prefrontal cortex in schizophrenia: I. regional cerebral blood flow evidence. *Archives of general psychiatry, 43*, 114–124.
Weiskrantz, L., Elliott, J., & Darlington, C. (1971). Preliminary observations on tickling oneself. *Nature, 230*, 598–599.
Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen, E. (2001). Autonomous mental development by robots and animals. *Science, 291*, 599–600.
Westheimer, G. (2008). Was helmholtz a bayesian? *Perception, 37*, 642–650.
Wetterich, C. (2020). *The probabilistic world*. arXiv preprint arXiv:2011.02867.
Wiese, W. (2014). Perceptual presence in the kuhnian-popperian bayesian brain. In *Open MIND. Open MIND*. Frankfurt am Main: MIND Group.
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review, 9*, 625–636.
Wilson, P., Humpston, C., & Nathan, R. (2021). Innovations in the psychopathology of schizophrenia: a primer for busy clinicians. BJPsych. *Advances*, 1–11.
Windt, J. M., & Noreika, V. (2011). How to integrate dreaming into a general theory of consciousness–a critical review of existing positions and suggestions for future research. *Consciousness and Cognition, 20*, 1091–1107.
Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in cognitive sciences, 5*, 487–494.
Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269*, 1880–1882.
Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural networks, 11*, 1317–1329.
Wolputte, S. V. (2004). Hang on to your self: Of bodies, embodiment, and selves. *Annu. Rev. Anthropol., 33*, 251–269.
World Health Organization. (2001). *The World Health Report 2001: Mental health: new understanding, new hope*. World Health Organization.

Yamashita, Y., & Tani, J. (2012). Spontaneous prediction error generation in schizophrenia. *PLoS One, 7*, e37843.
Yang, G. J., Murray, J. D., Repovs, G., Cole, M. W., Savic, A., Glasser, M. F., Pittenger, C., Krystal, J. H., Wang, X.-J., Pearlson, G. D., et al. (2014). Altered global brain signal in schizophrenia. *Proceedings of the National Academy of Sciences, 111*, 7438–7443.
Yoon, J. H., Maddock, R. J., Rokem, A., Silver, M. A., Minzenberg, M. J., Ragland, J. D., & Carter, C. S. (2010). Gaba concentration is reduced in visual cortex in schizophrenia and correlates with orientation-specific surround suppression. *Journal of Neuroscience, 30*, 3777–3781.
Zaadnoordijk, L., Besold, T.R., & Hunnius, S. (2019). A match does not make a sense: on the sufficiency of the comparator model for explaining the sense of agency. Neuroscience of consciousness, 2019, niz006.
Ziemann, U., Wittenberg, G. F., & Cohen, L. G. (2002). Stimulation-induced within-representation and across-representation plasticity in human motor cortex. *Journal of Neuroscience, 22*, 5563–5571.