Prerequisites for an Artificial Self

Verena V Hafner, Pontus Loviken, Antonio Pico Villalpando, Guido Schillaci

frontiers

Prerequisites for an Artificial Self

Verena V. Hafner[1*], **Pontus Loviken**[2,3], **Antonio Pico Villalpando**[1], **Guido Schillaci**[1,4]

[1] *Adaptive Systems Group, Computer Science Department, Humboldt-Universität zu Berlin, Germany*
[2] *Softbank Robotics, Paris, France*
[3] *University of Plymouth, UK*
[4] *The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy*

Correspondence*:
Verena V. Hafner
hafner@informatik.hu-berlin.de

ABSTRACT

Traditionally investigated in philosophy, body ownership and agency - two main components of the minimal self - have recently gained attention from other disciplines, such as brain, cognitive and behavioural sciences, and even robotics and artificial intelligence. In robotics, intuitive human interaction in natural and dynamic environments becomes more and more important, and requires skills such as self-other distinction and an understanding of agency effects. In a previous review article, we investigated studies on mechanisms for the development of motor and cognitive skills in robots (Schillaci et al., 2016). In this review article, we argue that these mechanisms also build the foundation for an understanding of an artificial self. In particular, we look at developmental processes of the minimal self in biological systems, transfer principles of those to the development of an artificial self, and suggest metrics for agency and body ownership in an artificial self.

1 INTRODUCTION

People can usually easily recognise their own body and the results of their own actions. This apparently simple skill likely contributes to what makes us feel as separate entities in the world (Van Den Bos and Jeannerod, 2002) and it is indeed fundamental for interacting with the environment and with other individuals. A current research trend suggests that the *minimal self* - the pre-reflective experience of being a self, or the awareness of oneself as a subject of experience (Blanke and Metzinger, 2009) - would be characterised by two important aspects: a sense of body ownership - *I feel corporal sensations as uniquely belonging to my own body* - and a sense of agency - *I feel being in control of my own actions* (Gallagher, 2000).

Topics such as body ownership and agency that have traditionally been investigated in philosophy have recently gained attention from other disciplines, such as brain, cognitive and behavioural sciences, and even robotics and artificial intelligence. Some neuroscientists, for example, interpret certain human mental disorders - such as schizophrenia - as the result of a disrupted sense of the self (Nelson et al., 2014; Klaver and Dijkerman, 2016; Frith et al., 2000; Sterzer et al., 2016). In robotics, intuitive human interaction in natural and dynamic environments becomes more and more important, and requires skills such as self-other distinction and an understanding of agency effects (Belpaeme et al., 2018; Holthaus and Wachsmuth, 2012).

Developmental psychologists study the emergence of self-awareness from very early stages of development. Self-awareness would unfold already during the first months of life, when infants seem to start having a sense of how their own body is situated in relation to other entities in the environment (Rochat, 2003). Infants at 5 months of age, for example, are able to distinguish their own leg movements from those of another infant, when they are displayed in a mirror (Rochat, 2003). These action-effects have been studied in infants using different modalities including sound (Paulus et al., 2012).

These findings represent a valuable source of inspiration for roboticists, whose aim is to develop autonomous robots capable of living in and interacting with the human society. Developmental robotics addresses this challenge by implementing methods and algorithms for motor and cognitive development in artificial systems inspired by infant development (Cangelosi and Schlesinger, 2015). In developmental robotics, state of the art machine learning techniques are applied to computational models, creating artificial systems that can adapt to new situations and learn in an open-ended fashion. The emergence of the self represents a key step in cognitive development. Therefore, there is a growing interest in the developmental robotics community on implementing processes capable of enabling the experience of the self - with phenomena such as sense of body ownership and agency - in artificial agents.

On the other side, robots can represent valuable tools to investigate phenomena of subjective experience typical of humans. In fact, robots are equipped with sensors and actuators that can be inspected and controlled during their operations. What the robot sees and perceives, and its internal states can be logged and further analyzed which is obviously not possible in humans. If robots were capable of detecting and recognising their own body and movements, their interaction with the environment and with people would be much more efficient and natural. However, the questions about which computational processes are needed to implement a primitive sense of body ownership and agency in robots, and of how the ontogenetic process of the individual shapes the development of the self, are still open.

This manuscript follows-up a previous review paper (Schillaci et al., 2016), in which we investigated studies on mechanisms for the development of motor and cognitive skills in robots. In this review paper, we argue that the same mechanisms also build the foundation for the development of an artificial self. In fact, in infants, the self seems to emerge along the motor and cognitive development of the individual (Lagercrantz and Changeux, 2009). Implementing similar processes in artificial systems may provide insights also in the possibility to develop an artificial self. In this work, we address the role of developmental processes in the emergence of an artificial self, and we suggest the concept of *self-manifolds* in artificial systems and the use of metrics for establishing the boundaries of an artificial self.

The review paper is structured as follows. First, in section 2, we revisit the concepts addressed in our previous review (Schillaci et al., 2016) and frame them within the context of the development of an artificial self. In particular, we present advances in the study of behavioural and computational components that allow autonomous motor and cognitive development in artificial systems. We discuss how these components can build the foundation for an artificial self. In order to do so, we ask whether and how the minimal self is affected during the ontogenetic process of the individual, and how open-ended learning and social interaction can shape the development of an artificial self, and then review robotic studies addressing this question. In section 3, we review studies on metrics and boundaries of the human self, and propose their use also for artificial systems. Finally, in section 4, we provide our conclusions and open challenges in the quest for the development of an artificial self.

2 BEHAVIOURAL AND COMPUTATIONAL COMPONENTS

In the robotics literature, the study on the artificial minimal self is young and fragmented. Unfortunately, a study presenting a comprehensive overview on the robotic investigations on this topic is missing. Nonetheless, many articles can be found providing interesting insights on aspects and prerequisites that can be related to the development of an artificial self. Two recent papers highlight both aspects of the human minimal self and an artificial minimal self. Georgie et al. (2019) look at developmental indices and behavioural measures of the minimal self, and Lanillos et al. (2019) look into computational models of neurological disorders related to the minimal self. In particular, they look into the balance between sensed and predicted sensory effects in ASD and schizophrenia.

In a previous review paper (Schillaci et al., 2016), we investigated studies on mechanisms for the development of motor and cognitive skills in robots. In particular, we identified three main behavioural and computational components that can enable autonomous acquisition of motor skills and the implementation of basic cognitive capabilities: (1) exploration behaviours; (2) internal body representations; (3) sensorimotor simulations. In this review, we extend the review provided in Schillaci et al. (2016) by creating links to the topic of the development of an artificial self, beside introducing more recent robotic studies on related topics. We particularly focus on those ones that propose strategies to scale up with motor and cognitive development. We extend exploration behaviours with artificial curiosity and sensorimotor simulations with predictive processes in order to strengthen the aspects of the development of a minimal self. All three components are processes or cognitive skills that run in parallel and independently from each other and can be seen as building blocks of the minimal self as discussed later.

2.1 Self-exploration behaviours and artificial curiosity

Human fetuses seem to already have some limited control on their body, as they react to touch, sound, smell, and pain, and even show facial expressions responding to external stimuli (Lowery et al., 2007). Some researchers (Lagercrantz and Changeux, 2009), though, believe that these reactions may have subcortical nonconscious origin and that, only shortly after birth, newborns show signs of basic self-awareness. In fact, developmental studies provide evidence about infant behaviours displaying some level of self-awareness in their first weeks of life (Rochat, 2011). Nonetheless, whether - and to what extend - self-awareness is present at birth, developmental researchers believe that it would unfold during early stages of development (see Rochat (2003) for empirical evidence and proposals). However, why and how self-awareness exactly would emerge during infancy are still open questions and in particular there are no thorough theories or computational models explaining their function. Hart and Scassellati (2011) argue that self-identification algorithms are the first step towards a more comprehensive model of the robotic self.

There is a general consensus on recognising the important role in the development of self-awareness to the perceptual experiences that toddlers undergo when exploring and playing with their surroundings. The self would emerge through the active interaction with one's physical and social environment (Verschoor and Hommel, 2017). Indeed, exploration behaviours are recognised as the means for motor and cognitive development in infants, as well as in robots (see Schillaci et al. (2016) for a review). Several studies investigate the cognitive mechanisms and drives behind exploration and play in infancy. In infants, curiosity - which is usually inferred through their use of prolonged visual attention to stimuli (Benson and Haith, 2010, pag.157-167) (Grgič et al., 2016) - is thought to drive the emergence of ordered developmental trajectories, including in domains such as vocal development, imitation and tool use discovery (Oudeyer, 2018; Acevedo-Valle et al., 2018). This is contrary to earlier belief that infants learn by random actions, but rather that their actions are goal-directed from the very start (Von Hofsten, 2004).

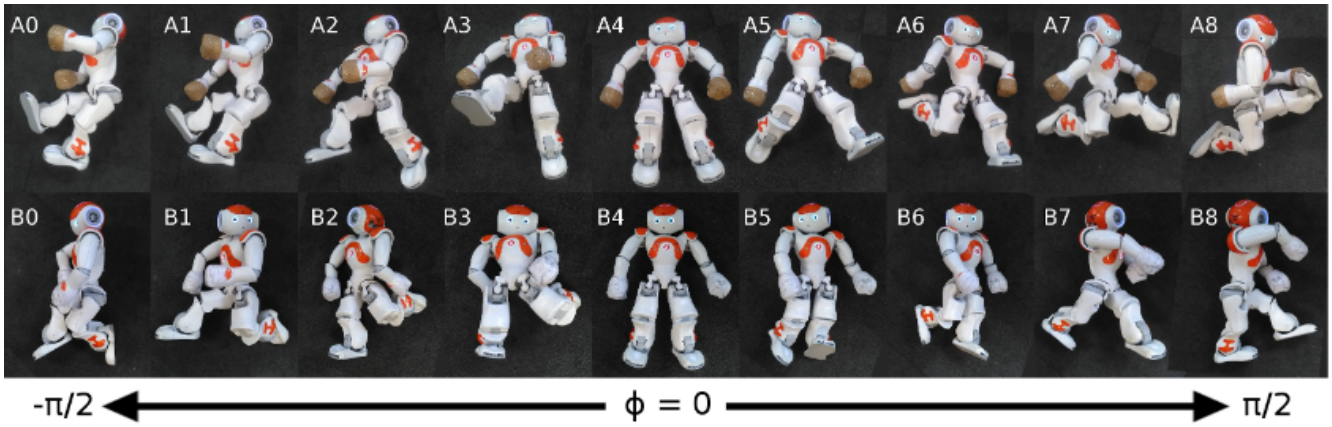

Figure 1. Curiosity-based learning method for humanoid robots using postures and regions. This image shows an example of postures learned after 30 minutes of online learning (Loviken et al., 2018). A and B represent two independent runs, and the number indicate the state. Each state is responsible for an interval of angle ϕ, where ϕ is the torso's orientation in relation to the ground. A demonstration video can be found at this URL: `https://www.youtube.com/watch?v=QzZsJxyGGIk`).

Infants' curiosity, play and exploration - and the likely goal-directed nature of their actions - have attracted the interest of developmental roboticists. In fact, studies on artificial curiosity have demonstrated how mechanisms for goal-directed exploration can be used to efficiently learn robot dynamics, even if the artificial system is characterised by complex high-dimensional embodiments. Artificial curiosity goes beyond novelty detection that would drive the agent to novel, but not necessarily predictable regions of its sensorimotor space. In contrast, artificial curiosity drives the agent towards regions where the learning progress can be maximised (Oudeyer et al., 2007). The main difference to typical machine learning scenarios is that the agent creates its own training samples for a desired learning trajectory.

The first studies on artificial curiosity and exploration in robots were limited, in a way. Although promising and demonstrating that curiosity-driven and exploration behaviours can efficiently solve inverse and forward kinematics problems, they mostly focused on relatively simple tasks, such as reaching actions for robot manipulators. Prolonged and incremental learning, until recently, was not a main priority in these studies. Indeed, it is still a great challenge in the whole robotics community. Seemingly, assuming that, in infants, self-awareness is a result of complex and prolonged interactions and experiences, the study on the development of an artificial self has to address, as well, how self-awareness would unfold along incremental learning in robots.

Recently, interesting studies have been published on topics close to this line of thoughts. For instance, studies in the literature on goal-directed exploration in artificial systems proposed ways to scale up learning to multiple task spaces (Forestier and Oudeyer, 2016; Forestier et al., 2017) or to domains where exploration of a task space requires action planning in multiple steps (Loviken and Hemion, 2017; Loviken et al., 2018). Figure 1 shows the results of a curiosity-based learning method for humanoid robots, where the sensory space was partitioned into a disjoint set of finite elements. In this space, every element was seen as an independent goal-babbling problem and a planning module could be added by observing transitions between the different elements (Loviken and Hemion, 2017; Loviken et al., 2018).

Acevedo-Valle et al. (2018) studied intrinsic motivation systems in the context of early vocal development which further develop through social reinforcement. An artificial agent was endowed with a proprioceptive mechanism, which was used to prevent the execution of unreachable motor configurations or invalid

(painful) configurations. Moreover, the authors introduced an expert instructor which produced correct utterances whenever the exploring autonomous learner was emitting similar (although still not correct) sounds. This resulted in a social reinforcement, which provided clues to the learner of interesting sensorimotor regions to explore.

Interesting advances have been made also in the context of goal generation. For instance, Mannella et al. (2018) show how an artificial system can autonomously generate goals to be used in an intrinsic motivation system to explore and to gather knowledge about its own body. In Schillaci et al. (2020), the authors present an architecture for curiosity-driven goal-directed exploration behaviours on a camera-equipped robot arm. A combination of deep neural networks for offline unsupervised learning of low-dimensional features from images, and of online learning of shallow neural networks was used. The artificial curiosity system assigned interest values to a set of pre-defined goals, and drove the exploration towards those that were expected to maximise the learning progress. Moreover, the authors proposed the integration of an episodic memory system to face catastrophic forgetting issues, typically experienced when performing online updates of artificial neural networks. The results showed that adopting an episodic memory system not only prevented the computational models from quickly forgetting knowledge that have been previously acquired, but also provided new avenues for modulating the balance between plasticity and stability of the models.

In humans, the self develops along the ontogenetic process of the individual. This is closely related to mechanisms of open-ended learning and social interaction, but also on the establishment and refinement of plastic body representations. The next section will provide an overview of recent studies on body representations in artificial systems.

2.2 Body representations

Many researchers have suggested theories in trying to explain the experience of body ownership and agency, and self-awareness in general. Sense of agency and sense of body ownership seem to be strongly linked, but many empirical studies still investigate them separately from each other. The appearance of the first signs of self-awareness in newborns seems to be dependent to the establishment of thalamocortical connections (Lagercrantz and Changeux, 2009). In general, the sense of body ownership seems to be strongly intertwined with an internal representation of the body maintained by our brain. Here we adopt the conceptual clarification by Gallagher (1986) between *body image* and *body schema*, where body image is a conscious representation or image of the body, whereas body schema is a non-conscious representation of sensorimotor skills. While we interact with the environment, we generate a rich set of multi-modal sensory and motor experience (Schillaci et al., 2016). This information has been proposed to be integrated in a sort of a body schema into our brain, which would keep an up-to-date representation of the positions and configurations of the different body parts in space (Maravita et al., 2003; Hoffmann et al., 2010). Moreover, the body schema very likely undergoes a continuous process of adaptation, as humans and animals follow an ontogenetic process where corporal dimensions and morphology change over time. The way in which we represent and feel our body seems to strongly rely on these representations, which would integrate inputs from different sensory modalities (Azañón et al., 2016). Scientists carried out experiments to explore how the brain combines information from the flow of sensory input data to create a feeling of body ownership, such as the famous experiment of the rubber hand illusion, where the participant is confused by the sight of a fake hand and synchronised sensory stimulation (Botvinick and Cohen, 1998).

Some researchers in cognitive development link the construction of the self to the experience encoded in a sort of autobiographical memory (Nelson, 2003). Pointeau and Dominey (2017) review a range of robotic experiments that address different aspects of the self and relate them to the definition of the self

as given by Neisser (1995). Ulric Neisser proposed five types of self-knowledge that correspond to five distinct components of the self: ecological, interpersonal, conceptual, temporally extended, and private. The *ecological self*, that is "the individual situated in and acting upon the immediate physical environment" (Neisser, 1995), is perhaps the level which is most interesting here, and it is rather easy, given the current robot technologies, to design robotic experiments addressing it. Ecological proprioception is integrated with different modalities of sensory information concerning one's own body as interacting within the environment (Gallagher, 2007). The tactile modality has received particular interest from researchers on subjective experiences, and on their impairments in patients with brain disorders. Van Stralen et al. (2011), for instance, studied how self-touch influences the structural representation of one's own body and found that self-touch may be modulating impairments in body ownership.

Developmental roboticists have also focused their attention onto the role of the tactile modality in the formation and maintenance of body representations. For instance, Zenha et al. (2018) studied how a body schema can be adapted incrementally in a humanoid robot based on touch events. Hoffmann (2017) studied the role of self-touch experiences in the formation of a self. Self-touch would provide redundant information that would facilitate the formation of a body representation. Timing and synchrony has been identified also as an important feature in support to the integration of information from multiple modalities within a body representations. Nabeshima et al. (2005) present a robotic study in support of that.

Hoffmann et al. (2018) studied a self-organising model for body representation on an iCub humanoid robot with an artificial pressure-sensitive skin. In particular, the proposed framework was used to learn a topographic representation of the robot's body surface from experience, that is by receiving tactile stimulations all over its artificial skin, including multi-touch stimulations.

2.3 Sensorimotor Simulations and Predictive processes

A growing number of scientists now consider the brain as an active organ of inference (Picard and Friston, 2014; De Ridder et al., 2013; Kirchhoff, 2018). Self-awareness and self recognition are thought to be dependent also on predictive processes - or sensorimotor simulations - implemented by the brain (Hohwy, 2013; Apps and Tsakiris, 2014; Friston, 2018). Predictive processes may have several functions, but one important is that of sensory attenuation. Pyasik et al. (2019) showed that felt ownership of a fake hand in the rubber hand illusion experiment caused attenuation of somatosensory stimuli generated by its movements comparable to the attenuation of self-generated stimuli. Burin et al. (2018) also investigated the influence of timing on the effect of agency.

Similar computational models can be implemented into robots to provide them with predictive capabilities. Sensorimotor predictions and prediction errors can be recorded and analysed, as well. In humans – in contrast – such properties cannot directly be observed and controlled. Bechtle et al. (2016) and Lang et al. (2018) implemented internal models into a humanoid robot to study how body representations can emerge from sensorimotor experience, and how predictive processes can be run through these computational tools. They found that prediction errors can serve as a cue to distinguish between self-generated perceptual events and those generated by other subjects. Moreover, they showed how predictive processes can be used to attenuate self-body perception (see figure 2). Lang et al. (2018) adopted a convolutional neural network for implementing a forward model, which generates image predictions from low-dimensional proprioceptive and motor states (see figure 3).

Pico et al. (2016) demonstrated that a two-wheeled mobile robot was capable of detecting unexpected changes in the environment and able to classify motor behaviours by comparing the ego-noise generated by its motors with the ego-noise prediction of its internal model. In a first experiment, several ego-noise

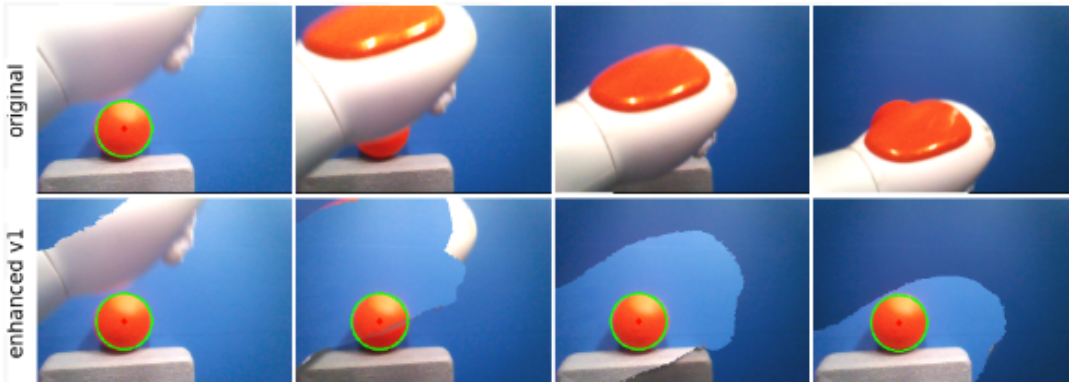

Figure 2. Self-body attenuation through predictive processes (Lang et al., 2018). A humanoid robot Nao is moving its arm in front of an object. The first row shows the frames recorded from its camera. The second row shows the enhanced frames, where self-body perception is attenuated. The attenuation is aided by a forward model, which anticipates the pixels where the robot arm will be visualised, after executing an intended motor command.

.

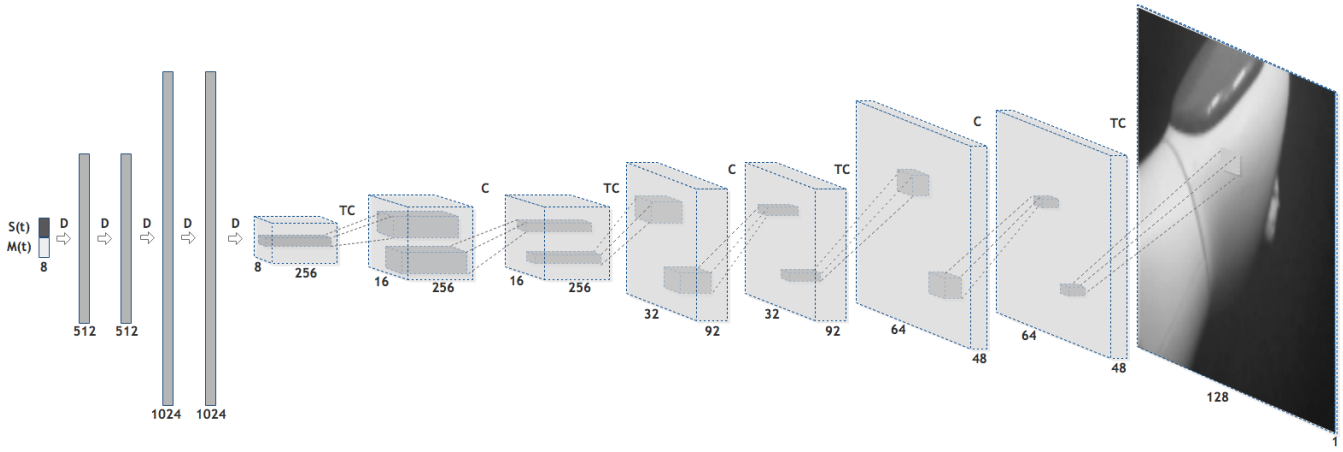

Figure 3. An illustration of the forward model adopted in Lang et al. (2018) for generating image predictions from low-dimensional proprioceptive and motor states through a convolutional neural network. Legend: S(t): sensory state at time t. M(t): Motor command sent at time t. D: Dense, i.e. fully connected, neural network layer. C: Convolutional neural network layer. TC: Transposed Convolutional neural network layer. Every layer except the last (output) one is followed by a ReLU activation unit (not shown) (Lang et al., 2018).

.

prediction models have been trained, each of them with a different motor command pattern. All models were then fed with a particular motor sequence, obtaining a series of ego-noise predictions. The robot was able to determine the correct motor command pattern by selecting the model with the lowest ego-noise prediction error. In a second experiment, one ego-noise forward model has been trained by implementing random motor babbling on the robot in a flat arena. The model was tested by making the robot do a series of runs from side to side of the arena while calculating ego-noise predictions. A ramp was then added in the middle and the runs were repeated. A comparison between the ego-noise prediction errors generated in the flat arena and those of the arena with the ramp on the middle, showed that the ego-noise prediction error increased when the robot was over the ramp. This demonstrated that the robot was able to detect changes in the inclination of the surface it moves only by making ego-noise predictions.

Predictive models can also be used for robot imitation. Pico et al. (2017) utilised robot ego-noise as a mean for communicating intended actions among robots. In an experiment, a robot generated a series of ego-noise audio (emulated by a loudspeaker) representing an intended motor command sequence and conveyed it to another robot. The receiver robot obtained auditory features from the ego-noise through a convolutional autoencoder. These audio features were then fed into an inverse model in order to obtain motor command predictions, which were similar to the motor commands that generated the audio produced by the sender robot.

Winfield (2018) describes a range of different experiments with artificial agents running internal simulations of themselves, others, and the environment, and compares these skills to an artificial Theory of Mind. ”Theory of mind is the term given by philosophers and psychologists for the ability to form a predictive model of self and others” Winfield (2018). These internal simulations show how to increase robot safety (Blum et al., 2018) by anticipating self and other behaviour (Winfield and Hafner, 2018).

Predictive processes have also been studied by Hinz et al. (2018) in the context of the rubber hand illusion. The authors analysed the drift in the perception of the real limb towards the fake limb, which would suggest an update of body estimation resulting from stimulation. In particular, they compared body limb drifting patterns of human participants with the end-effector estimation displacement of a multisensory robotic arm enabled with predictive processing perception. They observed similar drifting patterns in both human and robot experiments, suggesting that the perceptual drift is due to prediction error fusion, rather than hypothesis selection.

Touch seems to be a more direct sense, which could be trusted more for prediction than distant senses such as vision. It also equally concerns sense of agency and sense of body ownership. Ciaunica (2017) emphasizes the developmental aspects of touch, self-touch and intersubjective touch. An interesting aspect of predicting the sensory consequences of touch is the feeling of ticklishness, that has been addressed by Sarah Blakemore in a paper with the title ”Why can’t you tickle yourself” (Blakemore et al., 2000). This phenomenon of ticklishness has also been shown in mice recently (Ishiyama and Brecht, 2016). In a preliminary study on touch prediction in artificial systems, Stiehler and Hafner (2017) could show how a predictive model learns to predict the sensory consequences of touch. The sensory consequences of self-touch are usually more predictable than those of being touched by someone else. The sensation of ticklishness might be triggered by specific changes in prediction error over time, but there is little work so far on this topic. Quantitative studies showed that self-generated forces are perceived in the tactile modality as weaker than externally generated forces of the same magnitude, suggesting again that sensory consequences of a movement are anticipated and attenuated (Shergill et al., 2003).

Vicente et al. (2016) showed how predictive process can also support adaptation of body schemas. The authors combined predictions made by a learned internal model with the actual visual feedback to improve the perceptual skill of a humanoid robot.

The aforementioned studies suggest that predictive processes - as simulations of sensorimotor activities - are important tools for implementing basic cognitive capabilities in artificial systems, and may represent necessary building blocks for providing robots with subjective experiences, such as those typical of the minimal self.

3 METRICS FOR AN ARTIFICIAL SELF

As mentioned before, the minimal self is often described by two major building blocks: a sense of body ownership and a sense of agency. Both are subjective measures (articulated by the word ’sense’), and

can vary between individuals, over time, and depending on the situation. As has been shown in various experiments, for example in the rubber hand illusion (Botvinick and Cohen, 1998), and in virtual reality studies (Banakou et al., 2018; Blanke and Metzinger, 2009), both the sense of body ownership and the sense of agency can be altered in humans. This points towards a certain plasticity of the brain's body representation. Predictive capabilities play a major role in maintaining a consistent minimal self. Based on our self-models, we as humans anticipate the effects of our own actions and can thus monitor them. Longo et al. (2008) for example take a psychometric approach to the question of embodiment and sense of agency.

In artificial agents, a similar measure for a sense of body ownership and a sense of agency could be identified. As discussed in the previous sections, most models related to agency and ownership rely on forward models and internal simulations, and have permanent access to a prediction error. When such a model is embodied in an artificial agent, the agent has also direct access to this measure. Michel et al. (2004), for instance, showed in a robotics study that extensions of the self in the visual field can be identified by learning the time delay between actions and their effects.

What could be the necessary requirements of measuring self-ness in artificial agents? In analogy to prediction and anticipation in the human minimal self, a sense of agency and a sense of body ownership should be linked to changes in the prediction error in artificial agents over time as well. Preliminarily ignoring the complex dynamics of the prediction error, the lower the error in the prediction of the consequences of self-generated actions, the stronger a sense of agency and body ownership.

Given the considerations taken above, we can define a self-manifold in sensorimotor space with the following properties: It is dynamic, as it can change with body growth and the acquisition of new skills; it is adaptive, where the error tolerance can vary according to the specific context and the states of the system and of the surrounding environment.

The self-manifold outlines the boundaries of the self, both related to body ownership and agency, which cannot be clearly separated. A concrete example of learning manifolds in sensorimotor space, however not related to the concept of self, can be found in Laflaquière et al. (2015). The boundaries of the self related to body ownership are closely related to notions of peripersonal space (PPS) (Clery and Hamed, 2018). The same can hold for agency if we consider multisensory channels including tactile information and assume temporal and cross-modal predictions (Clery and Hamed, 2018).

Prediction errors - such as those produced by forward models - may be used for determining the boundaries of the self-manifold in the sensorimotor space of artificial agents. Hereby, we encourage further robotics investigation within this research line, as it may provide insights in the understanding of the human self and in the implementation of the artificial self.

This idea follows the argument of Gallese and Sinigaglia (2010) who envision the bodily self as a manifold of action possibilities that cannot be reduced to any form of proprioceptive awareness. Including the action possibilities necessarily needs a system able to make predictions about the consequences of own actions. Actions not only include physical body movements and change of postures, but also interaction with the external world, including interaction with objects but also other agents (see Neisser (1995)'s notion of interpersonal self).

For simplification, we only consider prediction errors caused by actions affecting the peripersonal space of the agent. A self-metric for an artificial agent is a systematic way to assign a value to each suitable instance of an agent self. It should allow us to compare the self-ness of one agent at a certain instant in time to the self-ness of another agent or the same agent at another instant in time.

Nonetheless, there are still open issues that need to be solved for deciding on such a metric: how big is the time window for normalisation and what other timing issues arise; what are the modalities to include or exclude; and which are suitable computational models for multimodal integration. Such a metric will also allow to decide the balance of predicted information versus perceived information and might ultimately shed light on mechanisms of disturbances of the self in humans.

Similarities to the concept of the self-manifold can be found with that of the *markov blanket* (Kirchhoff et al., 2018). Organisms tend to self-organise within a coherent whole, maintaining a boundary that separates their internal states from the external world. A markov blanket has been theoretised as defining the boundaries of such systems in a statistical sense. If taking the theoretical standpoint of the Free Energy Principle, as proposed by Friston (2013) , this would mean that organisms maintain their integrity by minimising variational free energy (surprise) over their internal states. That is, they maximise evidence for their own models, i.e. their own existence (Kirchhoff et al., 2018). In predictive coding, free energy is associated with prediction errors. The free-energy bound, or markov blanket, can be associated with a prediction error boundary. A self-manifold may thus be formalised as a markov blanket around the sensorimotor states of an agent.

4 CONCLUSIONS

In this manuscript, we studied the literature on developmental processes for an artificial self. We reviewed a number of works addressing the self in artificial systems and suggesting basic behavioural and computational components that may serve for the implementation of subjective experiences in robots. However, many questions and challenges in the development of an artificial self still remain open.

In section 2, we reviewed the behavioural and computational components necessary to develop an artificial self - inspired by models of the human self - in the three areas ”Self-exploration behaviours and artificial curiosity”, ”Body representations”, and ”Sensorimotor simulations and predictive processes”. These ingredients of an artificial self have been studied extensively in robotics and computational modeling, and will need to be integrated for a full understanding of the self using computational methods.

A common trend in both analytic sciences such as psychology and neuroscience and synthetic sciences such as robotics is to look more into the developmental processes that shape the self. This allows us to identify prerequisites and test existing theories of the self.

In section 3, we pointed out that beside the challenging task of implementing such mechanisms in artificial systems, there is a need for defining and designing metrics for an artificial self. We suggested requirements for such a self-metric and identified properties of a self-manifold as being adaptive and dynamic. Although we are far from establishing whether artificial agents can ever undergo subjective experiences, these metrics may provide support and insights in the investigation of the self, in both robots and humans.

To conclude this review, we suggest a number of open challenges of the artificial self. In particular, there is a need of integrating the three main behavioural and computational components mentioned above: Self-exploration behaviours and artificial curiosity, body representations, and sensorimotor simulations and predictive processes.

Moreover, further investigation is required in addressing the following overall challenges: designing models for multimodal integration in lifelong learning robotics setups; working on a refinement of self-metrics; identifying difference and complementarity between agency and body ownership; realising

the integration of temporal and intentional binding effects within predictive computational models; and resolving synchronisation as well as conceptual issues.

In robotics, we can access internal states and inspect sensorimotor and prediction information. However, to what extent can this privileged point of view allow us to *state* - if ever possible - that a robot is undergoing subjective experience? Indeed, there is a need for further debating the possibility of phenomenological experience in artificial systems.

CONFLICT OF INTEREST STATEMENT

PL was employed by SoftBank Robotics and has received funding from the European Union's Horizon 2020 research and innovation programme APRIL. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

VVH and GS produced most of the text within this manuscript. PL and APV contributed to section 2, in particular discussing studies on goal-directed exploration (PL) and ego-noise representation and imitation (APV).

FUNDING

The work of GS, VVH and APV was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 773875 (EU-H2020 ROMI, Robotics for Microfarms) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 402790442 ("Prerequisites for the Development of an Artificial Self").

PL has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 674868 (APRIL), where VH is also an associate partner.

The work of GS has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 838861 (Predictive Robots)

REFERENCES

Acevedo-Valle, J. M., Hafner, V. V., and Angulo, C. (2018). Social reinforcement in artificial prelinguistic development: A study using intrinsically motivated exploration architectures. *IEEE Transactions on Cognitive and Developmental Systems* doi:10.1109/TCDS.2018.2883249

Apps, M. A. and Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews* 41, 85–97

Azañón, E., Tamè, L., Maravita, A., Linkenauger, S. A., Ferrè, E. R., Tajadura-Jiménez, A., et al. (2016). Multimodal contributions to body representation. *Multisensory Research* 29, 635–661

Banakou, D., Kishore, S., and Slater, M. (2018). Virtually being einstein results in an improvement in cognitive task performance and a decrease in age bias. *Frontiers in Psychology* 9, 917. doi:10.3389/fpsyg.2018.00917

Bechtle, S., Schillaci, G., and Hafner, V. V. (2016). On the sense of agency and of object permanence in robots. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (IEEE), 166–171

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. *Science Robotics* 3. doi:10.1126/scirobotics.aat5954

Benson, J. B. and Haith, M. M. (2010). *Language, memory, and cognition in infancy and early childhood* (Academic Press)

Blakemore, S.-J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *Neuroreport* 11, R11–R16

Blanke, O. and Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in cognitive sciences* 13, 7–13

Blum, C., Winfield, A. F. T., and Hafner, V. V. (2018). Simulation-based internal models for safer robots. *Frontiers in Robotics and AI* 4, 74. doi:10.3389/frobt.2017.00074

Botvinick, M. and Cohen, J. (1998). Rubber hands 'feel'touch that eyes see. *Nature* 391, 756

Burin, D., Pyasik, M., Ronga, I., Cavallo, M., Salatino, A., and Pia, L. (2018). "as long as that is my hand, that willed action is mine": Timing of agency triggered by body ownership. *Consciousness and Cognition* 58, 186 – 192. doi:https://doi.org/10.1016/j.concog.2017.12.005

Cangelosi, A. and Schlesinger, M. (2015). *Developmental robotics: From babies to robots* (MIT Press)

Ciaunica, A. (2017). The 'meeting of bodies': Empathy and basic forms of shared experiences. *Topoi* doi:10.1007/s11245-017-9500-x

Clery, J. and Hamed, S. B. (2018). Frontier of self and impact prediction. *Frontiers in Psychology* 9, 1073. doi:10.3389/fpsyg.2018.01073

De Ridder, D., Vanneste, S., et al. (2013). The predictive brain and the "free will" illusion. *Frontiers in psychology* 4, 131

Forestier, S., Mollard, Y., and Oudeyer, P.-Y. (2017). Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*

Forestier, S. and Oudeyer, P.-Y. (2016). Modular active curiosity-driven discovery of tool use. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on* (IEEE), 3965–3972

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface* 10, 20130475

Friston, K. (2018). Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). *Frontiers in psychology* 9

Frith, C. D., Blakemore, S.-J., and Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. *Brain Research Reviews* 31, 357–363

Gallagher, S. (1986). Body image and body schema: A conceptual clarification. *The Journal of Mind and Behavior* 7, 541–554

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences* 4, 14–21

Gallagher, S. (2007). Bodily self-awareness and object perception. *Theoria et historia scientiarum* 7, 53–68

Gallese, V. and Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia* 48, 746–755

Georgie, Y. K., Schillaci, G., and Hafner, V. V. (2019). An interdisciplinary overview of developmental indices and behavioral measures of the minimal self. In *Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2019, Oslo, Norway, August 19-22, 2019*. 129–136. doi:10.1109/DEVLRN.2019.8850703

Grgič, R. G., Crespi, S. A., and de'Sperati, C. (2016). Assessing self-awareness through gaze agency. *PloS one* 11, e0164682

Hart, J. W. and Scassellati, B. (2011). Robotic models of self. In *Metareasoning: Thinking about Thinking*, eds. M. T. Cox and A. Rajat (MIT Press)

Hinz, N.-A., Lanillos, P., Mueller, H., and Cheng, G. (2018). Drifting perceptual patterns suggest prediction errors fusion rather than hypothesis selection: replicating the rubber-hand illusion on a robot. *arXiv preprint arXiv:1806.06809*

Hoffmann, M. (2017). The role of self-touch experience in the formation of the self. *arXiv preprint arXiv:1712.07843*

Hoffmann, M., Marques, H., Arieta, A., Sumioka, H., Lungarella, M., and Pfeifer, R. (2010). Body schema in robotics: a review. *IEEE Transactions on Autonomous Mental Development* 2, 304–324

Hoffmann, M., Straka, Z., Farkaš, I., Vavrečka, M., and Metta, G. (2018). Robotic homunculus: Learning of artificial skin representation in a humanoid robot motivated by primary somatosensory cortex. *IEEE Transactions on Cognitive and Developmental Systems* 10, 163–176

Hohwy, J. (2013). *The predictive mind* (Oxford University Press)

Holthaus, P. and Wachsmuth, S. (2012). Active peripersonal spaces for more intuitive hri. In *IEEE-RAS 12th International Conference on Humanoid Robots*. 508–513. doi:10.1109/HUMANOIDS.2012.6651567

Ishiyama, S. and Brecht, M. (2016). Neural correlates of ticklishness in the rat somatosensory cortex. *Science* 354, 757–760

Kirchhoff, M. (2018). Predictive brains and embodied, enactive cognition: an introduction to the special issue. *Synthese* 195, 2355–2366. doi:10.1007/s11229-017-1534-5

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface* 15, 20170792

Klaver, M. and Dijkerman, H. C. (2016). Bodily experience in schizophrenia: factors underlying a disturbed sense of body ownership. *Frontiers in human neuroscience* 10, 305

Laflaquière, A., O'Regan, J. K., Argentieri, S., Gas, B., and Terekhov, A. V. (2015). Learning agent's spatial configuration from sensorimotor invariants. *Robotics and Autonomous Systems* 71, 49 – 59. doi:https://doi.org/10.1016/j.robot.2015.01.003. Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence

Lagercrantz, H. and Changeux, J.-P. (2009). The emergence of human consciousness: from fetal to neonatal life. *Pediatric research* 65, 255

Lang, C., Schillaci, G., and Hafner, V. V. (2018). A deep convolutional neural network model for sense of agency and object permanence in robots. In *8th Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (IEEE), 260–265

Lanillos, P., Oliva, D., Philippsen, A., Yamashita, Y., Nagai, Y., and Cheng, G. (2019). A review on neural network models of schizophrenia and autism spectrum disorder. *Neural Networks* 122. doi:10.1016/j.neunet.2019.10.014

Longo, M. R., Schuur, F., Kammers, M., Tsakiris, M., and Haggard, P. (2008). What is embodiment? a psychometric approach. *Cognition 107 (3)* , 978–998

Loviken, P. and Hemion, N. (2017). Online-learning and planning in high dimensions with finite element goal babbling. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 247–254

Loviken, P., Hemion, N., Laflaquière, A., Spranger, M., and Cangelosi, A. (2018). Online learning of body orientation control on a humanoid robot using finite element goal babbling. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 4091–4098

Lowery, C. L., Hardman, M. P., Manning, N., Clancy, B., Hall, R. W., and Anand, K. (2007). Neurodevelopmental changes of fetal pain. In *Seminars in perinatology* (Elsevier), vol. 31, 275–282

Mannella, F., Somogyi, E., Jacquey, L., O'Regan, K., Baldassarre, G., et al. (2018). Know your body through intrinsic goals. *Frontiers in neurorobotics* 12, 30

Maravita, A., Spence, C., and Driver, J. (2003). Multisensory integration and the body schema: close to hand and within reach. *Current biology* 13, R531–R539

Michel, P., Gold, K., and Scassellati, B. (2004). Motion-based robotic self-recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), vol. 3, 2763–2768

Nabeshima, C., Lungarella, M., and Kuniyoshi, Y. (2005). Timing-based model of body schema adaptation and its role in perception and tool use: A robot case study. In *Proceedings. The 4th International Conference on Development and Learning, 2005* (IEEE), 7–12

Neisser, U. (1995). Criteria for an ecological self. In *Advances in psychology, 112. The self in infancy: Theory and research*, ed. P. Rochat (North-Holland/Elsevier Science Publishers). 17–34

Nelson, B., Parnas, J., and Sass, L. A. (2014). Disturbance of minimal self (ipseity) in schizophrenia: clarification and current status. *Schizophrenia bulletin* 40. doi:10.1093/schbul/sbu034

Nelson, K. (2003). Self and social functions: Individual autobiographical memory and collective narrative. *Memory* 11, 125–136

Oudeyer, P.-Y. (2018). Computational theories of curiosity-driven learning. In *The New Science of Curiosity*, ed. G. Gordon (NOVA). 43–72

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation* 11, 265–286. doi:10.1109/TEVC.2006.890271

Paulus, M., Hunnius, S., Van Elk, M., and Bekkering, H. (2012). How learning to shake a rattle affects 8-month-old infants' perception of the rattle's sound: electrophysiological evidence for action-effect binding in infancy. *Developmental Cognitive Neuroscience* 2, 90–96

Picard, F. and Friston, K. (2014). Predictions, perception, and a sense of self. *Neurology* , 10–1212

Pico, A., Schillaci, G., Hafner, V. V., and 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)*. 246–251. doi:10.1109/DEVLRN.2016.7846826

Pico, A., Schillaci, G., Hafner, V. V., and Lara, B. (2017). On robots imitating movements through motor noise prediction. In *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 318–323. doi:10.1109/DEVLRN.2017.8329824

Pointeau, G. and Dominey, P. F. (2017). The role of autobiographical memory in the development of a robot self. *Frontiers in neurorobotics* 11, 27

Pyasik, M., Salatino, A., Burin, D., Berti, A., Ricci, R., and Pia, L. (2019). Shared neurocognitive mechanisms of attenuating self-touch and illusory self-touch. *Social Cognitive and Affective Neuroscience* 14, 119–127. doi:10.1093/scan/nsz002

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and cognition* 12, 717–731

Rochat, P. (2011). What is it like to be a newborn? In *The Oxford Handbook of the Self*, ed. S. Gallagher (Oxford University Press). 57–79

519 Schillaci, G., Hafner, V. V., and Lara, B. (2016). Exploration behaviors, body representations, and
520 simulation processes for the development of cognition in artificial agents. *Frontiers in Robotics and AI*
521 3, 39
522 Schillaci, G., Villalpando, A. P., Hafner, V. V., Hanappe, P., Colliaux, D., and Wintz, T. (2020).
523 Intrinsic motivation and episodic memories for robot exploration of high-dimensional sensory spaces.
524 *arXiv:2001.01982 [cs.AI]*
525 Shergill, S. S., Bays, P. M., Frith, C. D., and Wolpert, D. M. (2003). Two eyes for an eye: the neuroscience
526 of force escalation. *Science* 301, 187–187
527 Sterzer, P., Mishara, A., Voss, M., and Heinz, A. (2016). Thought insertion as a self-disturbance: An
528 integration of predictive coding and phenomenological approaches. *Frontiers in Human Neuroscience*
529 10
530 Stiehler, F. and Hafner, V. V. (2017). Touch prediction and development of the self. In *poster presentation*
531 *at the Workshop on the Development of the Self, at the 7th Joint IEEE International Conference on*
532 *Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)* (IEEE)
533 Van Den Bos, E. and Jeannerod, M. (2002). Sense of body and sense of action both contribute to
534 self-recognition. *Cognition* 85, 177–187
535 Van Stralen, H., Van Zandvoort, M., and Dijkerman, H. (2011). The role of self-touch in somatosensory and
536 body representation disorders after stroke. *Philosophical Transactions of the Royal Society of London B:*
537 *Biological Sciences* 366, 3142–3152
538 Verschoor, S. A. and Hommel, B. (2017). Self-by-doing: The role of action for self-acquisition. *Social*
539 *Cognition* 35, 127–145
540 Vicente, P., Jamone, L., and Bernardino, A. (2016). Online body schema adaptation based on internal
541 mental simulation and multisensory feedback. *Frontiers in Robotics and AI* 3, 7
542 Von Hofsten, C. (2004). An action perspective on motor development. *Trends in cognitive sciences* 8,
543 266–272
544 Winfield, A. F. T. (2018). Experiments in artificial theory of mind: From safety to story-telling. *Frontiers*
545 *in Robotics and AI* 5, 75. doi:10.3389/frobt.2018.00075
546 Winfield, A. F. T. and Hafner, V. V. (2018). Anticipation in robotics. In *Handbook of Anticipation:*
547 *Theoretical and Applied Aspects of the Use of Future in Decision Making*, ed. R. Poli (Cham: Springer
548 International Publishing). 1–30. doi:10.1007/978-3-319-31737-3_73-1
549 Zenha, R., Vicente, P., Jamone, L., and Bernardino, A. (2018). Incremental adaptation of a robot body
550 schema based on touch events. In *8th Joint IEEE International Conference on Development and Learning*
551 *and Epigenetic Robotics*. 119–124