White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247957/

Version: Published Version

Article:
Manneschi, L., Jimenez-Rodriguez, A., Camilleri, D. et al. (2026) Scaffolded acquisition of a robot self–world distinction in a layered architecture model of the developing self. Philosophical Transactions of the Royal Society B: Biological Sciences, 381 (1962). 20250153. ISSN: 0962-8436

https://doi.org/10.1098/rstb.2025.0153

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/rstb

CC BY

Research

Cite this article: Manneschi L, Jimenez-Rodriguez A, Camilleri D, Prescott TJ. 2026 Scaffolded acquisition of a robot self–world distinction in a layered architecture model of the developing self. *Phil. Trans. R. Soc. B* **381**: 20250153.
https://doi.org/10.1098/rstb.2025.0153

Received: 26 August 2025
Accepted: 5 May 2026

One contribution of 17 to a discussion meeting issue 'From being to identity: analysis and synthesis of the self'.

Subject Areas:
computational biology, cognition, neuroscience

Keywords:
cognitive neurorobotics, embodied cognition, anoetic self, sense of ownership, sense of agency, latent representation, neural ODEs, multisensory integration, superior colliculus, parietal cortex

Author for correspondence:
Tony J. Prescott
e-mail: t.j.prescott@sheffield.ac.uk

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8721458.

THE ROYAL SOCIETY PUBLISHING

Scaffolded acquisition of a robot self–world distinction in a layered architecture model of the developing self

Luca Manneschi[1], Alejandro Jimenez-Rodriguez[2], Daniel Camilleri[3] and Tony J. Prescott[1]

[1]School of Computer Science, The University of Sheffield, Sheffield, UK
[2]Department of Computing, Sheffield Hallam University, Sheffield, UK
[3]Kinematic Trees Ltd, Sheffield, UK

TJP, 0000-0003-4927-5390

How does the sense of self emerge in human development, and could an artificial system acquire a similar sense using algorithmic principles and given a suitable embodiment and experience? Here, we describe a machine learning framework, inspired by developmental psychology and neurobiology, that acquires a self–world distinction and minimal self-model. Unlike previous approaches, we introduce a gradual formulation, integrating correlation-based and predictive techniques within a layered architecture. Using a simulated humanoid, we show first how a self–world distinction can emerge through correlation-based self-supervision. This model uses dependencies between proprioception and vision to distinguish its own body from its surroundings during infant-like motor babbling. Next, we construct a model of how these sensory signals will evolve in time. This predictive capacity uses the self–world distinction acquired by the lower layer as input to neural ordinary differential equation models that learn continuous dynamics. Finally, we demonstrate scaffolding of the higher layer by the lower one, showing that predictive learning occurs more rapidly with the lower layer already in place. By integrating these mechanisms, our approach offers a framework for characterizing a minimal robotic self that can inform our understanding of the emergence of the sense of self in the human mind/brain.

This article is part of the discussion meeting issue 'From being to identity: analysis and synthesis of the self'.

1. Introduction

The *self–world distinction* is the separation of experience (perceptual and cognitive) as deriving from the self ('me') or from its counterpart, the world ('not me') [1]. For the human self, this distinction seems closely tied to the *sense of ownership* of the physical body and of signals emanating from inside body (or as relating to the body, for instance, when you observe yourself) and to the *sense of agency*—the experience of owning and originating actions and their consequences [2]. Following Piaget [3], the distinction between self and world may also be seen as a prerequisite for a *self–other distinction* with respect to the social world [4,5], that is, the separation of the mental life of the self from that of other selves [1,6]. Although conceptually we can distinguish these different aspects of the sense of self, understanding how they stand with respect to each other and how each might emerge in both evolution and development remains mysterious. Here, we take a synthetic 'understanding through building' approach [7–9] to begin to untangle these interdependencies.

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

As we will explore in this article, the self–world distinction is not all-or-nothing and can manifest in different ways. It can be non-conceptual, or *anoetic*, where the self is experienced as distinct but without any insight or self-awareness [10,11]. On the other hand, it can also be reflective, or *noetic*, where the self can be a subject of thought and judgement [10,12]. Finally, it can be *autonoetic* where the self can be conceived of and experienced as existing across time [12–14]. These three aspects can be present alongside each other. The anoetic aspects may contribute a feeling of the self as motivated, integrated and bounded, while the noetic aspects become more prominent during self-reflection and autonoetic when remembering the past or imagining the future.

In psychiatric conditions, different aspects of the self–world distinction can become disturbed, such as in *asomatognosia* where a part of the body is experienced as not belonging to the self, or in *autoscopia* where a visual hallucination of the self is projected into the outside world [15]. Likewise, altered states of consciousness, such as those induced by meditation or psychoactive drugs, can lead to a partial dissolution of the self–world boundary [16–18]. Experimental devices, such as the rubber hand illusion, and technologies, such as virtual reality and robotic telepresence, can also alter the experience of the self–world boundary, such that we might experience ownership of an artificial hand [19] or have a feeling of the self as being outside the physical body [20,21]. This evidence points to the actively and continuously constructed nature of the experience of selfhood [22].

In developmental terms, construction of a self–world distinction likely begins with the infant's own body and with the interoceptive, proprioceptive and somatosensory signals it generates [5]. For instance, basic visceral neural circuits, such as those underlying heartbeat, breathing and swallowing, emerge by the second trimester of pregnancy [23] and could provide a backdrop of predictable internal sensory dynamics to which new kinds of signals can be associated as also belonging to the self [14,24]. Proprioceptive signals provide evidence to the developing brain of its own agency as they vary in correlation with internally generated motor signals [5,25,26]. Somatosensory signals emanating from the skin—the body–world boundary—afford different experiences when the unborn baby touches innervated and non-innervated regions of her body [27]. This may provide the developing fetus with one of the earliest indications of her separate and bounded nature.

By the time we are born, human experience likely includes a distinction between self and world, albeit of an anoetic nature [5,27,28]. This is indicated, for instance, by a touch on the cheek of a neonate which will trigger a rooting response except when the infant touches herself [29]. In the newborn, hearing and vision offer new means of learning about the self through *reafference*—observation of the sensory consequences of self-generated movements and sounds [5,30,31]. For instance, when awake and alert, and in a supported position, a newborn may engage in arm and hand movements, a portion of which will fall within her visual field [32]. These movements, which appear exploratory but not goal-directed and are sometimes referred to as hand or motor 'babbling' [33,34], provide a new means of distinguishing the bodily self and its actions. By three months, exploratory movements are largely replaced by coordinated visually guided reaching that appears more intentional [35] and, by around six months, by reach and grasp behaviours that are suitably adapted to context [36,37].

In this contribution, we are specifically concerned with the question of how reafference might contribute to the emergence of the self–world distinction. We take a synthetic approach [7,9] using machine learning and (simulated) robotic modelling to explore how visual reafference, as in the infant's observations of her own limb movements, might guide the emergence of a visual self–world distinction (VSWD). We have previously argued that a unified experience of self emerges from the activity of multiple self 'subsystems' in the mind/brain [8,22,38]. Moreover, building on early proposals from Damasio [39], Panksepp [11,40] and Northoff *et al.* [41], we have also argued that this 'self system' is organized in the form of a layered architecture [42,43], whereby some core aspects of self—principally those related to anoetic senses of agency and ownership—are available in young infants (0–3 months of age), while other aspects emerge more gradually during childhood [14,22]. For instance, while the newborn is thought to live 'in the moment', by age two, the developing child will begin to distinguish past from future. However, this is still only the beginning of the emergence of a sense of the self in time. Full autonoetic experience, including the capacity to mentally 'time travel' into the past or future, may only arise alongside the acquisition of autobiographical memory capacities by age 4 or later [13,14,44].

According to this view, postnatal experience, alongside brain maturation, is critical in assembling the human sense of self. Here, we propose and test, using a robot simulation of motor babbling, a layered architecture hypothesis for the acquisition of a VSWD during the first few months of infancy. Specifically, as illustrated in figure 1, we first model a learning layer, inspired by Hebbian principles, that uses the correlation between visual and proprioceptive signals to provide an initial decomposition of the visual world into self and non-self components. We then extend this model with a predictive learning layer that constructs a dynamical model of the relationship between proprioception and visual experience of the self and is able to predict how each could evolve in the near future. We show that by training the simpler correlation-learning layer first, and by providing its *latent* (internal) representation of self–world as input to the predictive layer, we are able to 'scaffold' [43,46] the capacity of the latter to learn effectively and quickly.

As further set out below, both learning systems are forms of deep neural network building on current approaches in machine learning. To be clear, we are not arguing that these are strongly analogous to specific brain processes since we are not modelling brain circuits directly or attempting to closely match the visual and proprioceptive systems of the infant. Nevertheless, we do consider that this work can be informative in understanding the organization and emergence of the self–world distinction in the human mind/brain and return to this topic in the discussion in relation to potential cortical and subcortical substrates for distinguishing self from non-self. We next describe our robot simulation environment and the design and testing of, first, the correlation learning model and, second, the full model with correlation-based and predictive learning layers.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

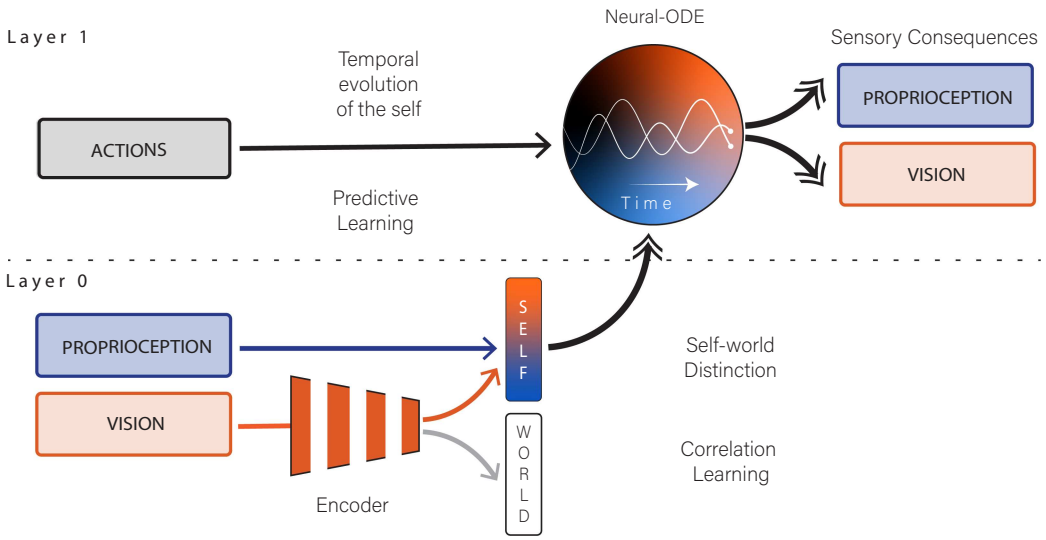


Figure 1. A layered architecture for the self–world distinction. Layer 0 acquires the ability to segregate sensory signals generated by self and world through correlation-based learning between proprioceptive and visual signals. This provides an anoetic sense of ownership over parts of the visual field occupied to the body. Layer 1, which becomes active once layer 0 training has taken place, adds a predictive capability that anticipates how both visual and proprioceptive signals will evolve in time. Sense of agency can be characterized as the experience of congruence between the predicted sensory consequences of action and the actual consequences [26,45]. Since layer 1 generates exactly these kinds of predictions, it could be viewed as providing the foundations for an anoetic sense of agency (see Discussion).

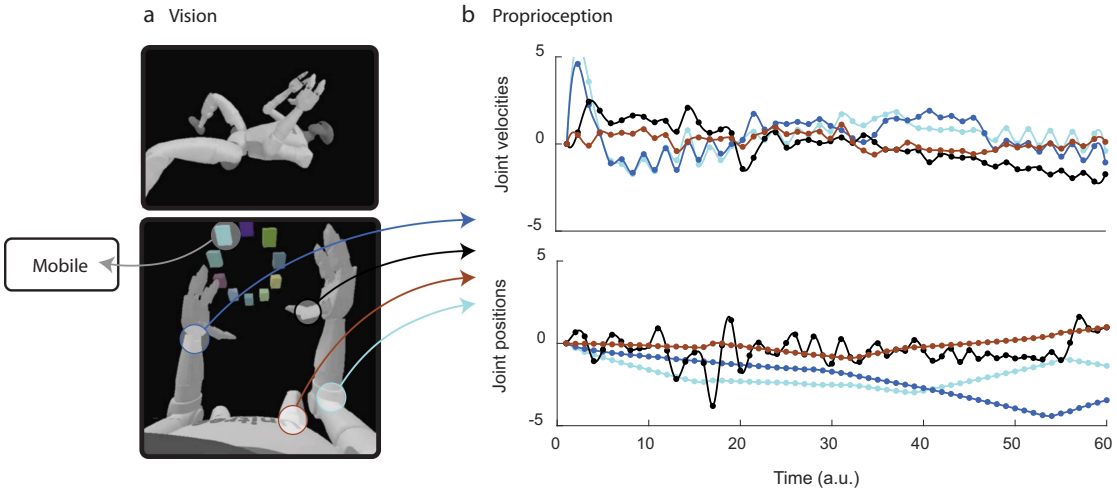


Figure 2. Illustration of the simulation set-up and data collection for the motor babbling task. (a) The simulated humanoid robot lies on its back (top view) observing its body, arms and hands through an RGB camera (bottom view) located in the head. We place a revolving visual pattern resembling a baby mobile in the robot's visual field against an otherwise black background. The mobile provides an external, moving object in the scene, which is not part of the robot, and can serve as a target for world segmentation. RGB images of 128 x 128 resolution are obtained at 30 fps. (b) Example proprioceptive data. Proprioceptive signals from 23 joints are collected at 20 Hz; the plots show example of how these evolve over time.

2. Layered architecture model

(a) Humanoid robot simulation

We use a full-physics simulation of the Unitree G1 humanoid (figure 2) using a faster than real-time GPU (graphics processing unit)-accelerated simulation, ManiSkill [47]. We select the G1 as it has a human-like morphology, including two legs, two arms and two three-fingered hands. The robot's 23 degrees of freedom allow it to make movements of the arms and legs that resemble those of humans, although it has no neck joints. The robot is equipped with a head-mounted monocular RGB-D (red, green, blue–depth) camera with segmentation, with 85.2°× 58°× 94° field of view. We use only the monocular two-dimensional RGB images for our study since infants' eyes are poorly coordinated and stereo vision does not emerge until three to four months [48]. We place the robot on its back to resemble a human infant in a cot and add to the visible environment a collection of coloured boxes arranged in a rotating circle that resembles a baby 'mobile' (a patterned toy often suspended above a baby's cot that rotates to provide enriched visual experience).

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

To generate training data for our models, we implement motor babbling by sampling target hand positions using a three-dimensional Gaussian with an assigned variance per dimension. Targets are represented as the desired change in position, $\Delta(x, y, z)$, with respect to the robot base, for each of the end-effectors (both hands). When a new target is chosen, we actuate the robot using Manikskill's built-in PID (proportional, integral, and derivative) controller, which issues expected delta movements to the robot's joints. This is done per end-effector, i.e. each hand samples from its own Gaussian distribution of target positions and is appropriately actuated to move towards it. Once the target position is reached, a new one is selected. The simulation is configured to run at 100 Hz (physics time step). Every five simulation steps (i.e. at a frequency of 20 Hz), the control loop receives sensory data—proprioceptive data (positions, velocities and effort per joint) and camera frames specified as 'rgb+depth+segmentation' with a resolution of 128 × 128—and issues control commands. The depth map is ignored, while the segmentation image, which provides object identities (ids) per pixel, is further processed to create a self-segmentation ground truth where all the ids related to the robot's links are given an id of 1. Additionally, for each processed frame, we compute a scalar per-pixel estimate of optic flow to provide a motion cue.

As will become clear in the following sections, our layered architecture learns to disentangle *self* and *world* in its internal representation via self-supervised learning, while also predicting the consequences of its actions. To assess the quality of the resulting latent spaces, and to guide optimization, we adopt decoders that reconstruct the input modalities, making explicit what information is captured by each latent. Ideally, the self latent should reconstruct the body, whereas the world latent should reconstruct the external object (the mobile). For visual reconstructions, we quantify performance using *pixel-wise accuracy* and *soft Dice coefficient* [49] measures computed over a region of interest (ROI). Full definitions are provided in the supplementary material; here, we summarize the essential information. Given a decoded attention map, we retain the top- k attended pixels and compare them against the corresponding ground-truth segmentation. Pixel-wise accuracy is computed after hard-thresholding the reconstruction: we retain the top- k pixels, binarize them and compare the resulting mask to the ground-truth segmentation. By contrast, the soft Dice coefficient is computed on the continuous reconstructed values within the ROI, yielding a graded measure of overlap between prediction and ground truth without binarization. We will report pixel-wise accuracy when assessing the behaviour of the proposed architecture and illustrating basic qualitative trends as it has a simple and intuitive interpretation. The soft Dice coefficient is widely used in the development of image segmentation models as it provides a graded measure that can capture finer-grained overlap information; it is therefore useful for comparing *between* alternative training methods. We use it here when comparing different model variants in §2d and in the supplementary material.

(b) Correlation-learning layer

(i) Overview

The correlation-learning layer, illustrated in figure 3a, is designed to differentiate between self and non-self, while also constructing a latent representation that can support higher-level components. We employ an encoder–decoder structure where the encoder is designed to separate self-related and world-related information into two distinct latent spaces. These representations enable the decoders to independently reconstruct both the agent (self) and its surrounding environment.

To reinforce the separation of self and world, the visual features contributing to the self-representation are passed through an information bottleneck, promoting a focus on self-related content. This can be understood as follows. Following the self pathway (black and red arrows in figure 3a), the latent representation associated with the self is used to reconstruct both proprioceptive and visual information. By contrast, the world-related pathway (grey arrows) is dedicated solely to reconstructing visual input. The demand for proprioceptive reconstruction in the self pathway acts as the bottleneck, effectively filtering for visual features that are specifically relevant to the self. Optimization proceeds in two stages. In stage 1 (10^4 iterations), we encourage the model to form a minimal sense of self by suppressing the world pathway. This forces the self pathway to learn an initial *spatial mask*, in the form of a 128 × 128 image filter with values in the range 0–1 per pixel, that will guide subsequent learning. In stage 2 (up to 10^5 iterations), we train all pathways jointly. The significance of this sequential training process will be further analysed in §2d.

Overall, this network can be regarded as a correlation-learning mechanism because the encoder–decoder design compresses the inputs into a latent space that must capture statistical dependencies, including cross-modal correlations, in order to accurately reconstruct the information at the current time. As such, future observations are not required; the model operates solely on information available at the present time and does not perform any explicit prediction. The current model processes visual and proprioceptive data; however, the architecture is inherently modular and could be extended to integrate additional sensory signals, such as audition or touch.

(ii) Methods

We next explain the network architecture in more detail. As shown in figure 3a, we adopt separate visual encoders and decoders for the self and the world to prevent information leakage between the two processing pathways. Each encoder/decoder is implemented as a convolutional neural network, while the decoder for proprioceptive information is a multi-layer perceptron (MLP). We denote the visual input and its reconstruction by $\mathbf{x}^v(t)$ and $\hat{\mathbf{x}}^v(t)$, respectively, where both are $128 \times 128 \times 4$ tensors of per-pixel RGB and optical-flow values. Similarly, we write the proprioceptive input and its reconstruction as $\mathbf{s}(t)$ and $\hat{\mathbf{s}}(t)$, respectively,

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

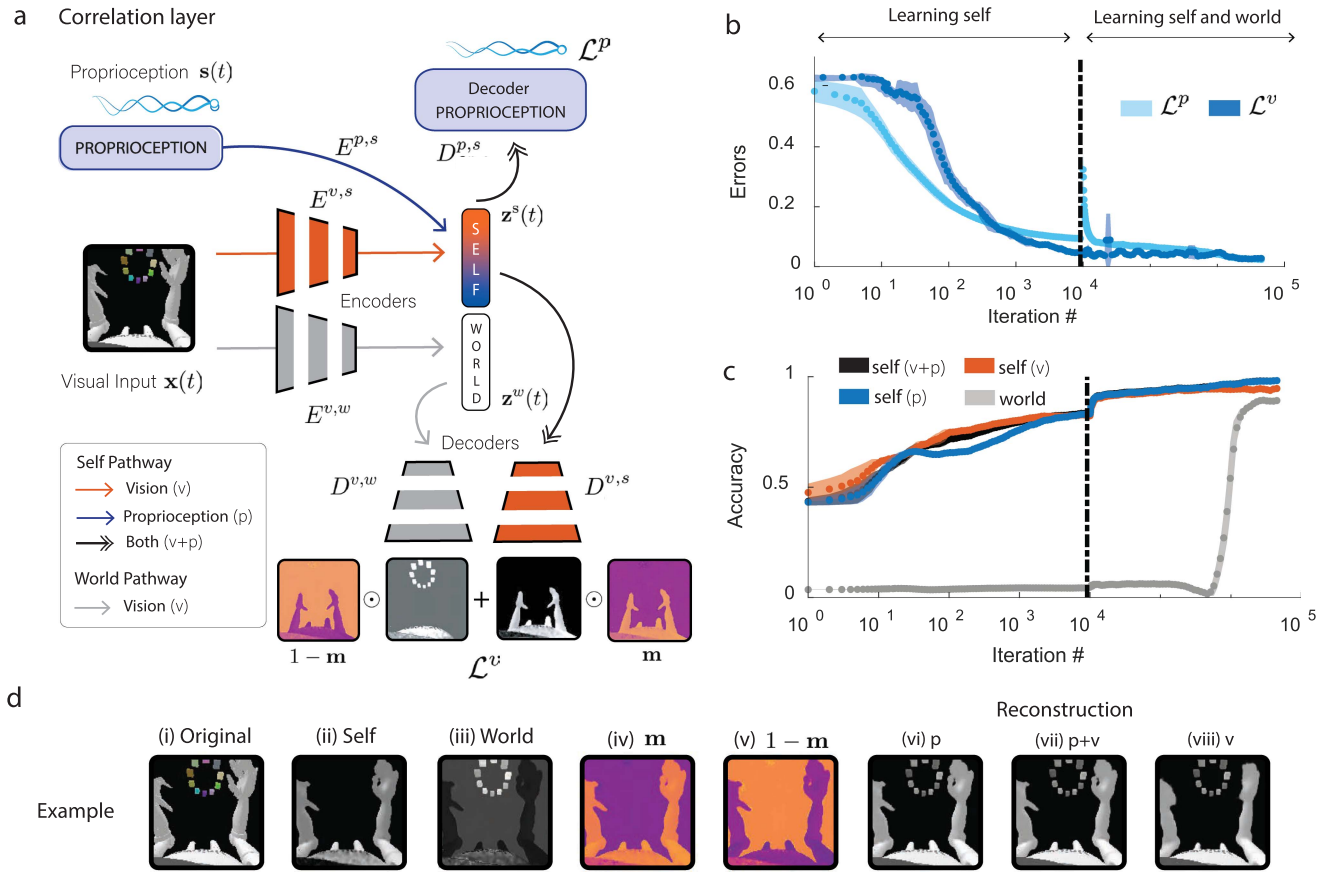


Figure 3. The correlation-learning layer. (a) Network architecture, where we adopt different latent spaces for the self and world information. (b) Learning curves for proprioceptive ($\mathcal{L}^p$) and visual ($\mathcal{L}^v$) reconstruction losses. The optimization proceeds in two stages, separated by the vertical dashed line at 10^4 iterations. In the first stage, the world pathway is suppressed and the self pathway acquires a basic self representation; in the second stage, both pathways are active, and the network shifts to jointly learning the self and the world. (c) Pixel-level segmentation accuracy corresponding to the input configurations *proprioception only (p)*, *vision only (v)*, *proprioception and vision combined (p+v)* and *world*. The curves are smoothed using a running average, and the shaded bands denote the running standard deviation, providing an indication of training stability. The *x*-axis follows a piece-wise log-scale to better visualize the two training phases. (d). Example reconstructions. From left to right: (i) the original image; (ii) the decoded image from the *self* latent, $D^{v,s}(\mathbf{z}^s)$; (iii) the decoded image from the *world* latent, $D^{v,w}(\mathbf{z}^w)$; (iv)–(v) the self and world spatial masks (lighter colours indicate higher attention); and (vi)–(viii) full reconstructions under three input conditions: *proprioception only (p)*, *combined model (p+v)* and *vision only (v)*. See text for further explanation.

where both are 46-dimensional vectors of joint positions and velocities. The self and world latent spaces are $\mathbf{z}^s(t)$ and $\mathbf{z}^w(t)$. Finally, we denote the visual self-, visual world- and proprioceptive encoders by $E^{v,s}_{\phi^{v,s}}$, $E^{v,w}_{\phi^{v,w}}$ and $E^{p,s}_{\phi^{p,s}}$, respectively, and the corresponding decoders as $D^{v,s}_{\theta^{v,s}}$, $D^{v,w}_{\theta^{v,w}}$ and $D^{p,s}_{\theta^{p,s}}$. For notational convenience, we omit the lowercase parameter indices of the encoder and decoder architectures from now on and implicitly assume that each function is defined by its own set of parameters.

Under this notation, the latent spaces are defined as

$$\mathbf{z}^s(t) = \alpha\, E^{v,s}\left(\mathbf{x}^v(t)\right) + (1-\alpha)\, E^{p,s}\left(\mathbf{s}(t)\right) \tag{2.1}$$

$$\text{and } \mathbf{z}^w(t) = E^{v,w}\left(\mathbf{x}^v(t)\right), \tag{2.2}$$

where $\alpha \in [0,1]$ balances the contributions of the visual and proprioceptive signals to the self latent and is randomly and uniformly sampled once per training sequence to promote robustness to changes in sensory-modality availability.

The proprioceptive reconstruction is $\hat{\mathbf{s}}(t) = D^{p,s}\left(\mathbf{z}^s(t)\right)$. For visual reconstruction, we combine the self and world decoders using a spatial attention mask $\mathbf{m}(t) \in [0,1]^{H\times W}$ (broadcast along channels):

$$\hat{\mathbf{x}}^v(t) = \mathbf{m}(t) \odot D^{v,s}\left(\mathbf{z}^s(t)\right) + \left(\mathbf{1} - \mathbf{m}(t)\right) \odot D^{v,w}\left(\mathbf{z}^w(t)\right), \tag{2.3}$$

where $\odot$ denotes the element-wise (Hadamard) multiplication. The spatial attention mask $\mathbf{m}(t) \in [0,1]^{H\times W}$ is computed from the decoded self-visual prediction as

$$\mathbf{m}(t) \;=\; f\left(\langle D^{v,s}(\mathbf{z}^s(t))\rangle_{\text{ch}} - c\right), \tag{2.4}$$

i.e. we first take a channel-wise average $\langle\cdot\rangle_{\text{ch}}$ to obtain a single $H \times W$ intensity map and then apply a shifted sigmoid $f(\cdot)$ (with shift c) to obtain values in [0,1]. The constant c controls the operating region of the sigmoid (in practice, any $c \in [0,1]$ works

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

well, and we set $c = 0$ in all simulations). Finally, the term $1 - \mathbf{m}(t)$ is the element-wise complement of $\mathbf{m}(t)$ and weights the world-related reconstruction accordingly.

Optimization of the layer is performed through backpropagation by minimizing the cost function:

$$\mathcal{L}_{\text{corr}} = \|\mathbf{x}^v(t) - \hat{\mathbf{x}}^v(t)\|_1 + ||\mathbf{s}(t) - \hat{\mathbf{s}}(t)||_2^2 + \beta \sum_{i=0}^{T-1} \|\mathbf{z}^s(t_i) - \mathbf{z}^s(t_{i+1})\|_2^2 = \mathcal{L}^v + \mathcal{L}^p + \beta \sum_{i=0}^{T-1} \|\mathbf{z}^s(t_i) - \mathbf{z}^s(t_{i+1})\|_2^2, \tag{2.5}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the mean absolute error and mean squared error, respectively, $t_i = t_0 + i\Delta t$ and β is a scaling factor of the temporal regularization term. This third factor in the loss function enforces temporal proximity across time in the latent space and preparing $\mathbf{z}^s$ to be modelled as a dynamical system (§2c). As noted above, there are two optimization stages. In stage 1, we set $\mathbf{z}^w(t) \equiv \mathbf{0}$ to completely suppress the world pathway; we then remove this constraint in stage 2. For further details regarding the architecture and specific hyperparameters, see the supplementary material.

(iii) Results

Figure 3b,c shows training curves for the layer 0 model, generated using the data described in §2a, while figure 3d provides examples of the visual data, the computed separation of self and world and reconstructed images for different input configurations. Each of these can be understood as follows.

Figure 3b reports the proprioceptive ($\mathcal{L}^p$) and visual ($\mathcal{L}^v$) reconstruction errors calculated on the test set as learning progresses (training iteration number on the x-axis). These show that errors in both domains decrease smoothly during stage 1, with a brief uptick at the transition to stage 2 when the world pathway becomes active and optimization enters a new regime.

Figure 3c shows pixel-level self/world segmentation accuracy, computed by matching the reconstructed self and world images to ground-truth segmentations obtained *a priori* for self and world, respectively (§2a). We plot three self-accuracy curves corresponding to testing the trained system with different input configurations: *proprioception only (p)*, *vision only (v)* and a *combination of the two (p+v)*; we also plot one curve corresponding to reconstruction of the visual *world*. These metrics are computed by evaluating the model on a test set at logarithmically increasing intervals and using ground-truth segmentation provided by the simulator (§2a). The plots show that, in stage 1, the self pathway learns a coarse segmentation of its own body against the environment, while the world pathway remains at chance level (since its output is suppressed). After the transition to stage 2, self accuracy exhibits a sharp jump and continues to improve, and the world pathway learns to focus on objects in the scene (the mobile), achieving strong final performance. A coherent sense of self should be resilient to the loss of a modality; this is demonstrated by the relatively small differences between the various *self* curves—all testing regimes perform well, though the small drop in the *no-proprioception (v)* condition visible at the end of training underlines the importance of proprioceptive input for an optimal self representation. We further notice that performance in the proprioception only (p) condition is comparable to the one attained in the combined (p+v) setting. The reason is that proprioception acts as an explicit information bottleneck used to induce a self–world distinction. Because the self latent is trained to reconstruct proprioceptive signals, the model learns a proprioception-centred representation and naturally relies on proprioceptive inputs whenever they are available. Visual input can still contribute, but it becomes most relevant when proprioception is weak or absent, where it provides robustness by allowing the self-state to be inferred from visual cues alone.

Finally, figure 3d provides qualitative examples of the capacity of the model to reconstruct the visual world and to decompose it into self and world components. As expected, the image reconstructed by the self decoder (d*ii*) concentrates on the agent's body, whereas the world decoder highlights objects in the scene (d*iii*), specifically, the mobile. Notably, the world representation exhibits lower luminance over the arms and hands, while retaining some information about the more static torso. This is consistent with the torso's limited motion and the fact that static, low-frequency information is easier to reconstruct, enabling the world pathway to represent some torso-related information despite primarily focusing on external objects. The masks (d*iv* and dv) partition the scene appropriately, and the three rightmost reconstructions (d*vi*–d*viii*) illustrate that all input configurations yield accurate results, with the combined (v+p) and proprioception-only (p) conditions performing best.

(c) Predictive-learning layer

(i) Overview

Beyond its role in generating a minimal VSWD, the correlation layer also supports the operation of the layer 1 predictive learning system. Specifically, its design encourages the emergence of a smoothly varying latent space that can be modelled as a dynamical system. After training, as described above, the weights in the correlation-learning layer are fixed. The predictive layer then builds on this system to provide a near-term forecast of what will happen next.

More specifically, the predictive layer is implemented as a neural ordinary differential equation (neural-ODE), i.e. an ordinary differential equation parametrized by a neural network. Rather than predicting the next system state directly, this network outputs the temporal derivative of the system state, and a numerical integrator is then used to obtain its evolution over time. The predictive layer thus allows us to evolve the latent state forward as a dynamical system driven by the planned action sequence. As a result, the model can imagine future latent states and, via the decoders, the corresponding future proprioceptive and visual observations. For comparison purposes, we also test a version in which the output of the model is provided as its own input

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

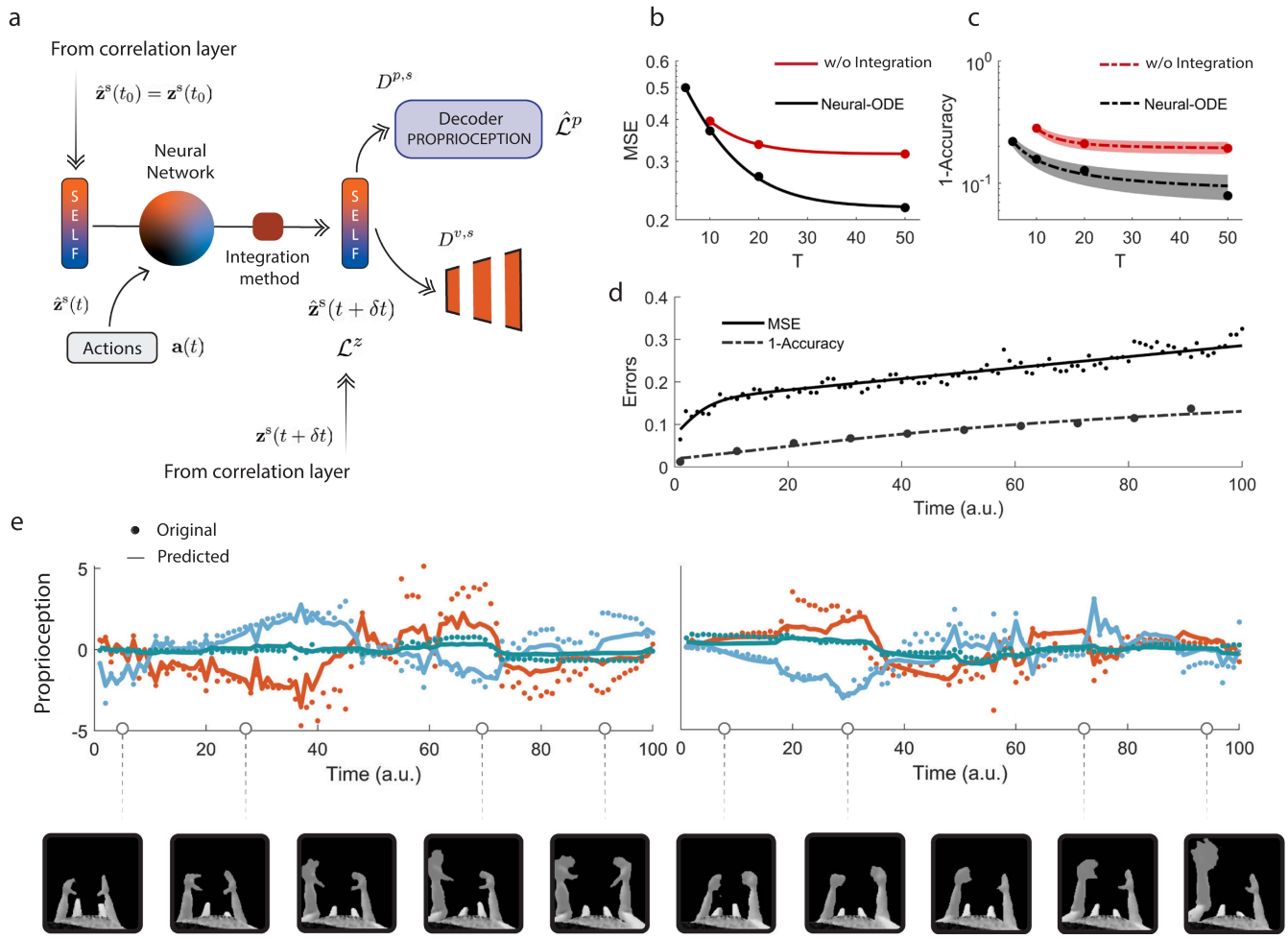


Figure 4. The prediction-learning layer. (a) Network architecture. (b) Mean squared error (MSE) of proprioceptive reconstructions for the neural-ODE and for an MLP trained without an explicit numerical integration scheme, as a function of the training sequence length (T). (c) As in (b), but reporting pixel-level error rate (1-accuracy). (d) Pixel-level accuracy and proprioceptive reconstruction error (MSE) as a function of prediction horizon for a neural-ODE model, with explicit integration, trained with $T = 50$. The plot illustrates the accumulation of error as the predictive layer rolls the latent state further into the future (time on the *x*-axis). (e) Two examples (left and right) of reconstructed proprioception signals (top) and exemplar visual predictions (bottom), obtained by decoding the forecasted self latent space. See text for further explanation.

in the next prediction step without a numerical integration procedure. This does not directly emulate a dynamical system but could generate similar outcomes. Optimization of this architecture proceeds with the *encoder* and *decoder* parameters kept frozen (unless otherwise specified in the experiments in §2d). This design lets the correlation layer scaffold the predictive layer and is also computationally efficient as explored further below.

To train the predictive model, we drive it with a sequence of external control signals corresponding to the relative change in hand position required to move to the next target position (as specified by the motor babbling); the outputs of the model are generated through the proprioceptive and visual decoders of the correlation layer for predictions of up to 100 time steps in the future.

(ii) Methods

As illustrated in figure 4a, the neural-ODE is defined over the self latent space $\mathbf{z}^s(t)$ and is driven by an external control signal $\mathbf{a}(t)$, corresponding to the actions the robot intends to perform. The purpose of this layer is to forecast the future latent representation and associated proprioceptive states, given an initial condition $\mathbf{z}^s(t_0)$ and a sequence of actions $\mathbf{a}(t)$. For the current model, the sequence of actions is the sequence of target hand positions (§2a).

Formally, letting $\hat{\mathbf{z}}^s(t)$ denote the predicted latent state, the dynamics of the neural-ODE are defined as:

$$\frac{\mathrm{d}\hat{\mathbf{z}}^s(t)}{\mathrm{d}t} = \mathbf{h}(\hat{\mathbf{z}}^s(t), \mathbf{a}(t)|\theta^{\mathrm{pred}}) \tag{2.6}$$

$$\hat{\mathbf{z}}^s(t_0) = \mathbf{z}^s(t_0), \tag{2.7}$$

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

where $\mathbf{h}$ is a neural network—a standard MLP in our implementation—parametrized by θ^{pred}, and the caret superscript $\hat{}$ makes explicit the difference between the modelled and target latent states. The loss function used to train the predictive component is:

$$\mathcal{L}_{\text{pred}} = \sum_{i=1}^{\mathrm{T}} \|\hat{\mathbf{z}}^{s}(t_i) - \mathbf{z}^{s}(t_i)\|_2^2 + \sum_{i=1}^{\mathrm{T}} ||\mathbf{s}(t_i) - D^{p,s}\left(\hat{\mathbf{z}}^{s}(t_i)\right)||_2^2 = \mathcal{L}^{z} + \mathcal{L}^{p}. \tag{2.8}$$

In equation (2.8), the first term aligns the latent trajectory predicted by the neural-ODE with the latent activities produced by the correlation layer, while the second term penalizes errors in reconstructing proprioception. During predictive training, we use the latent space consistency loss $\mathcal{L}^{z}$ in place of a visual reconstruction loss, thereby avoiding repeated and costly decoding passes. In all settings, unless otherwise stated, neural-ODE integration is performed using a second-order Runge–Kutta scheme with discretization step $\Delta t = 0.1$. To assess the importance of using an explicit numerical integrator and a neural-ODE formulation, we also evaluate an autoregressive MLP baseline that predicts the latent state forward in time directly, without any integration method. Further details of variant models and their computational costs are provided in §2d and in the supplementary material.

(iii) Results

In figure 4, we show how the mean squared error (MSE) of the proprioceptive prediction error (panel b) and the pixel-wise error rate (1–accuracy) (panel c) depend on the temporal horizon T (equation (2.8)) and on whether numerical integration is used. The reported metrics are averaged over 100 step-ahead predictions. Overall, the errors decrease as the model is trained to predict longer trajectories, and incorporating an integration scheme yields consistently better performance. These results highlight the importance of treating prediction as learning a dynamical process over extended horizons, rather than as a one-step forecasting problem. Figure 4d shows how the same measures (MSE of the prediction error and pixel-wise error rate) vary as a function of prediction horizon for a neural-ODE model, with explicit integration, trained with $\mathrm{T} = 50$. As expected, error accumulates as the predictive layer rolls the latent state further into the future; nevertheless, pixel-level accuracy remains close to 90% even 100 steps ahead. In other words, despite the accumulation of error, the model is able to generate plausible future trajectories.

Figure 4c presents two examples (left and right) of reconstructed proprioceptive information for three representative variables with a range of dynamic variation. Targets are shown as dots and predictions as solid lines together with the corresponding reconstructed visual inputs at representative time intervals (bottom images). Predictions are reasonably accurate for the first 20 time steps and then begin to show underestimation errors while typically in the right direction. However, even at longer prediction horizons, the images preserve a coherent representation of the body and arms, with only finer details gradually deteriorating. Note that we omit the world latent from this analysis because its dynamics are exogenous to the agent and not action-conditioned. Learning them would require interaction-aware and stochastic dynamics using, for instance, a neural stochastic differential equation (neural-SDE) model [50,51]; we leave this to future work.

(d) Scaffolding

(i) Overview/methods

Our architecture and learning procedures are deliberately modular—the model combines heterogeneous adaptive components with complementary cost terms, and optimization is performed in multiple stages that reflect an overall layered structure (figure 1).

In an earlier paper, Wilson & Prescott [43] defined *scaffolding* such that, given two processes, one with faster (earlier) dynamics, B, can serve as a scaffold for the organization of a slower (later) process, A, if the functional consequences of process B constrain A, such that some product of process A may be improved. These authors further suggested that scaffolding can occur without direct interaction between the two processes—all that is required is for process A to be measurably impacted by process B in a positive way.

In our architecture, we implement two forms of scaffolding motivated by our aim to understand the role of scaffolding in human development (as considered further in the Discussion). First, we pretrain the self pathway of the correlation layer (layer 0), while suppressing the world pathway, before simultaneously training both. This biases the model to find a way to reconstruct the visual scene using proprioceptive input rather than simply building a compressed representation of the scene. This is an example of scaffolding as the pretraining of the self pathway scaffolds learning in the world pathway even though they interact only directly (they share the same error signal and spatial mask). Second, we fix the encoders/decoders of the correlation layer prior to training the layer 1 predictive model. This stands as a more direct form of scaffolding as the self latent representation generated in layer 0 is provided as input to layer 1. Note that these two forms of scaffolding are similar in that both simplify the learning problem for the other part of the system. In the layer 0 case, this is by allowing the world pathway to focus on the non-self; in the layer 1 case, this is by providing as input a latent representation that is useful for learning how the self will evolve in time.

To test the utility of scaffolding, we next compare this *standard architecture* against variants in which scaffolding is removed or altered. From a machine learning perspective, it is also important to understand how scaffolding impacts on optimization. We therefore also test additional variants to better understand the behaviour of the scaffolded model as compared with end-to-end *backpropagation through time* (BPTT). End-to-end backpropagation is the standard approach for training deep neural networks as it enables gradients from the final objective to be propagated through all intermediate computations and, in recurrent settings,

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

through time, thereby updating all parameters jointly to directly minimize loss. Although we expect BPTT to have high accuracy, one of the benefits of the standard architecture is that it may be computationally more efficient than other variants particularly those that update all parameters throughout training. We explore this further below and in the supplementary material.

Here, we explore five variants as defined below. For completeness, we also investigate a number of further variants in the supplementary material. All variants use numerical integration in the predictive layer and are trained with a time horizon of $T = 50$ until convergence. With the exception of variant (i), all variants also employ the self-pretraining phase.

(i) *W/o layer 0 scaffolding*. We remove the self-pretraining of the correlation layer and train all layer 0 pathways jointly from scratch. This directly tests the utility of the first form of scaffolding described above.
(ii) *Continuous training*. We jointly optimize both layer 0 and layer 1 throughout the whole training process. This tests whether there is a benefit to training level 0 to convergence before training level 1, as compared to training both layers from the outset. In other words, this tests the utility of the second form of scaffolding described above.
(iii) *End-to-end BPTT*. We jointly optimize both layers and allow prediction errors to flow through the entire architecture so that the correlation layer parameters (including encoders and decoders) are updated using gradients from both equations (2.5) and (2.8). We expect this variant to have good performance but to be computationally costly.
(iv) *W/o proprioception*. The standard architecture but with both layers trained without proprioceptive input though still trying to predict future proprioceptive states (i.e. retaining the loss term $\mathcal{L}^p$). This condition tests whether visual information by itself is sufficient to induce a self–world distinction (for instance, through temporal correlation of moving body parts) and predict future self-states.
(v) *End-to-end backpropagation through time without proprioception (BPTT, w/o p)*. We jointly optimize both layers using backpropagation, as in variant (iii), but driving the system without proprioceptive input as in variant (iv). This condition again tests whether visual input is sufficient to induce a self–world distinction but this time under end-to-end optimization using backpropagation.

We should point out that it is also possible to continue to optimize layer 0 during layer 1 training, rather than freezing the layer 0 weights in the second training phase. This could further test whether having a stationary level 0 is beneficial to training level 1 (as compared to variant (ii) where both layers are trained from the outset). In practice, however, since layer 0 is trained to convergence during the initial phase, the weights would not change further during training of layer 1. In other words, this version will have a similar outcome to the standard model (although this would not be the case if the training data were to change).

(ii) Results

In figure 5, panel (a) shows exemplar image reconstructions from the self and world latent spaces for all five variants. Panels (b) and (c) report the standard model's soft Dice reconstruction scores and proprioceptive prediction errors, respectively, under different input regimes—*p*: proprioception only, *v*: vision only and *p*+*v*: a balanced mixture. The final two panels show the corresponding differences, post-convergence, between each variant and the standard baseline—panel (d) in terms of soft Dice scores and panel (e) for proprioceptive prediction errors. In these plots, self performance is reported at two horizons: $t = 0$ (lighter), which evaluates correlation-layer separability, and $t = 50$ (darker), which probes the predictive capacity of layer 1. In both plots, bars that extend to the left of the midline show outcomes that are worse than the standard model, while bars that extend to the right show better outcomes. Note that the soft Dice score for world reconstruction is provided for $t = 0$ only since we are not attempting to forecast future world states (§2c).

Compared to the standard model, the five variants behave as follows:

(i) *W/o layer 0 scaffolding*. Removing the level 0 pretraining causes the network to reconstruct the entire scene almost exclusively through the *world* pathway (panel a(i)). Intuitively, while the self pathway must learn cross-modal correlations and reconstruct both proprioception and vision, the world pathway can more rapidly discover a purely visual encoder–decoder (i.e. by performing a form of compression) that therefore dominates early training. The resulting model is rather crude and performs poorly on visual reconstruction (panel d), but the architecture is able to construct a self latent that is sufficient to predict future proprioceptive states (panel e).
(ii) *Continuous training*. This variant achieves similar MSE to the standard model (panel e), but exhibits reduced soft Dice when the system is driven by vision alone (panel d). A likely explanation is that, in this setting, the predictive layer is trained against targets produced by the correlation layer that continue to change during optimization. This can destabilize learning and increase susceptibility to suboptimal local minima. We further note that this approach is less computationally efficient than the standard model (see below).
(iii) *BPTT*. As expected, end-to-end optimization improves both soft Dice (panel d) and prediction MSE (panel e), but at a substantial computational cost since it requires BPTT and through the full architecture (see below).
(iv) *W/o proprioception*. Removing proprioception prevents the model from establishing a robust self–world distinction, leading to poor self and world reconstruction (panel a(iv)) compared to the standard architecture. The soft Dice scores are much lower (panel d) and MSE scores much higher (panel e) for all tests.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

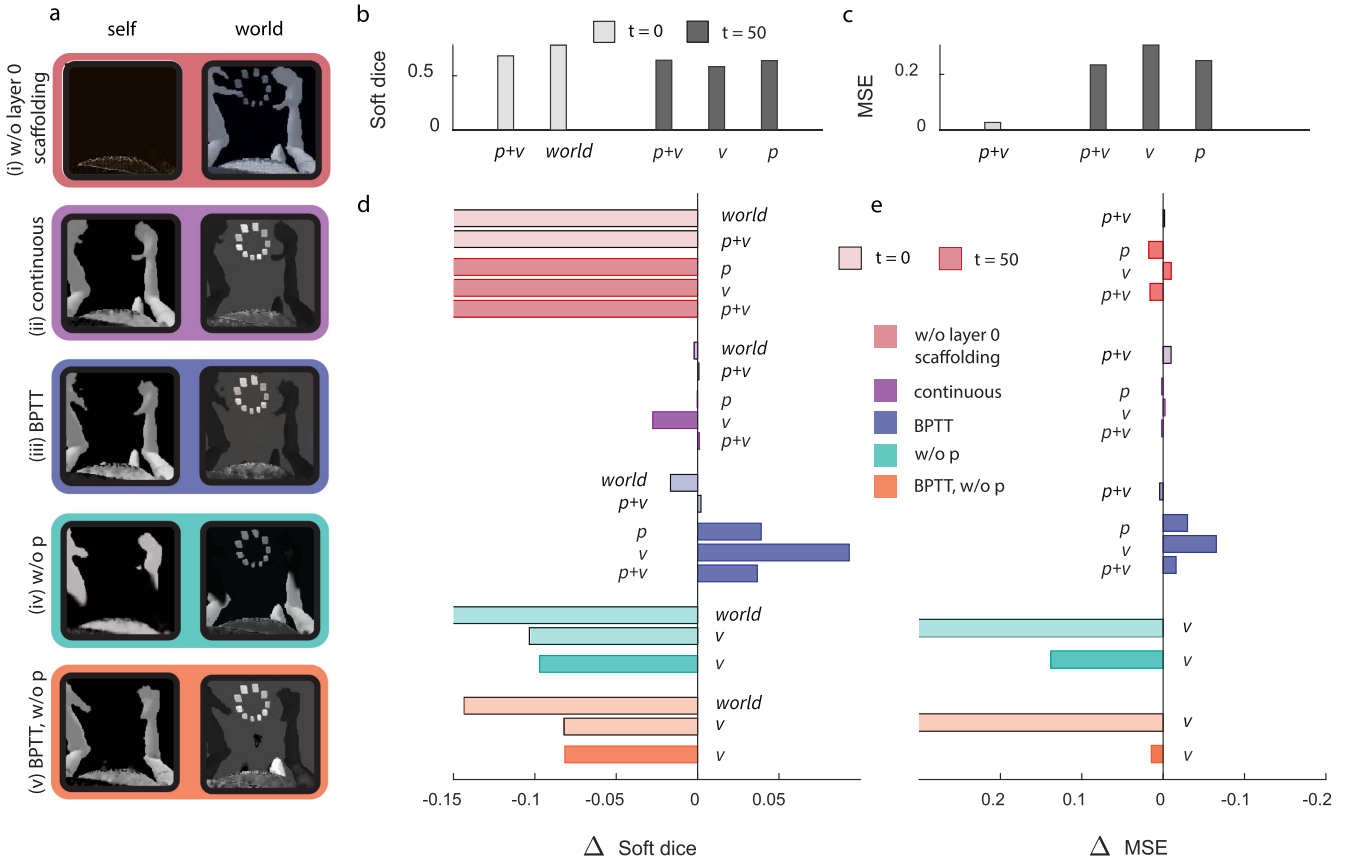


Figure 5. Results for model variants. (a) Example reconstructions (see figure 3d legend for description of elements), from top to bottom: w/o proprioception, BPTT w/o proprioception, continuous training, BPTT. (b) In lighter grey, self and world soft Dice for the reconstruction at the current time step ($t = 0$) and the standard model. In darker colours, self soft Dice when the system is driven by proprioception (p), vision (v) or a mixture of the two for predicting 50 steps in the future ($t = 50$). (c) Same as (b), but for the proprioception reconstruction error. (d) Accuracy differences with respect to the baseline standard model for the different settings explored (right is better). For variants (iv) and (v), trained without proprioception, we report performance under vision-only control, since the remaining settings require proprioceptive input. (e) Analogous to (d), but for the differences in the reconstruction errors.

(v) *BPTT, w/o p*. This variant has the same available information as variant (iv), which fails dramatically, but achieves marginally better performance. Indeed, it does surprisingly well at predicting future proprioceptive states (panel e). Overall, however, it fails to provide a satisfactory VSWD as indicated by the poor self–world separation in the image reconstructions (panel a(v)) and low soft Dice scores (panel d).

Overall, we observe that the staged training approach of the standard model performs well in these comparisons. Compared to variant (i), which lacks pretraining, it develops a reliable VSWD. Performance is similar to variant (ii) that allows joint optimization of both layers throughout the training process, and even marginally better on vision-only reconstruction of future self-states. Removal of the proprioceptive input (variant (iv)) demonstrates that the performance of the model leverages these internal signals in order to distinguish self from world and predict how the self will evolve.

As anticipated, end-to-end BPTT is able to improve on the results obtained with the layered architecture; however, this comes at some cost. By keeping the layer 0 encoders and decoders frozen during layer 1 training, the layered model obtains substantial efficiency savings—reducing the number of multiply–accumulate operations by around 98% compared to end-to-end BPTT. Indeed, we see substantial savings even in the continuous training setting (see the supplementary material for further explanation and detailed estimates). This performance improvement is achieved through the separation of timescales as predicted by Wilson & Prescott [43], with faster processes providing latent decompositions that scaffold acquisition of the slower predictive dynamics.

3. Discussion

With regard to the origins of the self in body exploration in early infancy, the developmental psychologist Phillipe Rochat wrote:

> Self-produced action causes the infant to have overlapping experiences through his other senses. [...] This experience specifies the body as differentiated from other objects in the environment. When my hand crosses my visual field, for example, I perceive that it is my hand and not someone else's, because I see it and feel it proprioceptively moving at exactly the same time and by a commensurate amount [5, p. 37].

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

In this contribution, we have investigated Rochat's proposal for the differentiation of self from world in infancy using a synthetic approach. Specifically, we have shown that a correlation-learning system can decompose the visual field of a simulated robot into self and non-self (figure 3). If proprioceptive signals are considered to already be part of the self because they originate from inside the body [5], then this learning system extends the self—in relation to an anoetic sense of body ownership—to parts of the visual world.

We have further shown that this minimal self–world distinction can usefully scaffold learning of continuous dynamical system models (neural-ODEs) that predict the future evolution of both proprioceptive and visual self-states (figure 4). Here, we have learned to predict the sensory outcomes of randomly generated 'babbling' actions; however, this predictive capacity should also be extendable to provide control signals for intentional, goal-directed actions (see below). Our model therefore aligns with theoretical discussions in which a pre-reflective sense of agency is seen as arising from congruence between the predicted sensory signals arising from voluntary action and the flow of reafferent sensory signals [26,45].

We next discuss our findings in relation to wider research on: (i) the minimal self in robotics, (ii) machine learning for multi-modal perception, (iii) the neural substrates for the human self–world distinction and its emergence during development, (iv) the role of scaffolding in the development of the human cognitive architecture, and (v) the future extension of the model towards noetic and autonoetic aspects of self.

(a) Cognitive neurorobotics approaches to understanding the minimal self

Within the broader field of synthetic science [52], the current work can be regarded as following a *cognitive neurorobotics* methodology that uses robot experiments to explore potential mechanisms underlying embodied cognition [53,54]. Rather than attempting to capture the detailed behaviour of neural circuits, this approach focuses on exploring high-level algorithmic principles and on aligning system dynamics and behaviour with findings from neuroscience and psychology; as such, it complements computational neuroscience modelling and neurorobotics approaches at lower levels of analysis [54]. By taking inspiration from the trajectory of human cognitive development, our study also aligns with wider research in the field of *developmental robotics* [55]. Here, we briefly review some related work that has taken a similar approach to understanding the self–world distinction and sense of agency through robotics. (For wider discussions of robotic models of self, see [14,22,56,57].)

In a pioneering study of robot sense of self, Tani [58] explored a layered architecture in which a recurrent network was trained to predict the sequence of perceptual outputs of an associative network controlling a vision-based mobile robot in a navigation task. Tani [58] suggested that a form of self emerges in the dynamics of the predictive network when its predictions diverge from the outputs of the lower-level controller. Tani & White [53] refined this proposal suggesting that a form of implicit self is already present when the robot correctly predicts its future sensory states (as in the definition of pre-reflective agency above), but that when this synchrony breaks down, and effort is required to refocus the system, the distinction between self and world is brought into consciousness. This proposal suggests one way in which the current architecture could be developed towards a noetic (reflective) sense of self, through monitoring of whether anoetic systems are achieving intended goals (see further discussion below).

An early study of robot self-recognition, by Michel *et al.* [59], used the expected time delay between action initiation and observed arm movement to judge whether visual motion was self-generated. Saegusa *et al.* [60] subsequently showed that correlation learning can be used to allow a motor babbling robot to visually recognise its own body parts. This was further demonstrated by Demirel *et al.* [61] who used correlation-based learning of the relationship between robot arm velocities and tracked visual tags. Bechtle *et al.* [62] trained a forward model to predict the visual consequences of a humanoid robot's own movements and used this to attenuate sensory signals arising from the robot's own body. These authors related the behaviour of their model system to a sense of agency and also argued that suppressing self-related visual signals could assist the robot in acquiring object permanence [3]. Lanillos *et al.* [63–65] have developed various models of robot self-perception based on predictive learning frameworks. These included a visual self–other distinction, generated by using proprioceptive signals to predict visual input, which allowed a robot to recognize its own mirror reflection [65]. Our model builds on these previous approaches by including both correlational and predictive learning into one architecture. Specifically, our correlation layer maps heterogeneous inputs (vision and proprioception) into a compact, well-posed latent space upon which predictive learning systems can be defined that forecast self-related states. This extends previous cognitive neurorobotics models by showing how a layered model can emulate aspects of the developmental timeline for sense of agency seen in human infants (see further below).

(b) Machine learning for multi-modal perception

The design of our model is also inspired by multi-modal state representation learning in robotics. For example, the proposed correlation layer is closely related to the variational autoencoder formulation of Meo & Lanillos [66]. This approach learns a joint latent from multiple sensors and enables cross-modal reconstruction under partial observability. Our architecture departs from this framework in three key ways. First, we explicitly factor the latent into a *self* component $\mathbf{z}^{s}$ and a *world* component $\mathbf{z}^{w}$, which (together) support information preservation while yielding an interpretable separation between agent and environment. Second, we endow the latent space with continuous-time dynamics via a neural-ODE driven by intended actions. This provides accurate prediction of future self-states through multi-step forecasting in latent space. Third, the presence of an explicit world representation, and of the dynamical prior, leads us to different architectural choices and loss terms from multi-modal

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

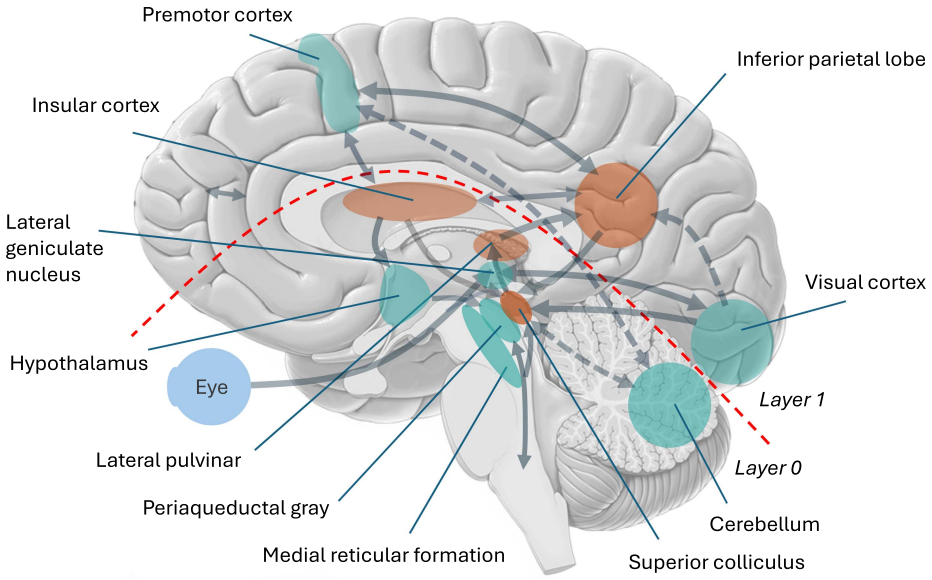


Figure 6. Some areas of the brain likely to be involved in the VSWD shown here above a sagittal section of the human brain. Key structures are highlighted in orange; other structures that process or relay signals relevant to a VSWD are highlighted in turquoise. Thick solid lines show direct connections; double arrowheads indicate that these are reciprocal; thick dashed lines show indirect connections. Only a subset of relevant structures and pathways are shown. Note that the insular cortex lies on an inner fold of the cortical sheet but is shown here overlying the sub-cortex. Layer 0 identifies a group of largely subcortical structures (but including insular) that are well-developed in the neonatal brain and could support an early form of VSWD similar to the correlation-based model in figure 3. We highlight the SC which receives direct input from the retina, and from multiple subcortical and cortical structures, and includes both a visual spatial map and activity related to proprioception. Layer 1 identifies later-developing cortical areas that could be involved in acquiring a more versatile form of VSWD similar to our predictive layer (figure 4). Of particular interest is the inferior parietal lobe (IPL), which is active during adult studies of agency and receives input from the SC via the lateral pulvinar and from insular cortex. Cerebellum is an important structure for learning to predict sensory signals; it may play an integrative role across the two layers. See text for further information.

autoencoder baselines. This connects our work to a growing literature in computer vision that has employed dynamical models to create image/video latent spaces. This research includes combining foundational neural-ODEs with latent ODEs [67,68], constructing second-order/physics-informed latent models [69,70] and neurally inspired or continuous-time video generation models [71,72]. These approaches have collectively demonstrated improved long-horizon rollout and more temporally consistent prediction compared to earlier work.

Predicting the future consequences of an agent's actions is closely related to the idea of using forward models in model-based reinforcement learning [73,74]. In this context, our architecture provides a simple self-supervised route to a self–world decomposition, which could help to structure learned world models by separating controllable (agent-related) from exogenous (environment-related) factors. Although speculative, such a factorization has clear potential advantages, including more modular reuse of predictive components across tasks and improved generalization when the agent or the environment changes independently.

(c) The layered brain architecture for self–world discrimination

Panksepp [11,40] described a 'primal self' as arising from the convergence of emotional (evaluative), sensory and motor schemata with a body map and a core motivational arbitrator situated in subcortical regions of the brain. Key structures relevant to this include the periaqueductal grey (PAG), the hypothalamus and the medial reticular formation, all of which are important motivational and emotional structures [11,75], and the superior colliculus (SC) a critical site for multi-sensory convergence that is also involved in regulating head, eye and body movements [76,77]. All of these structures are appropriately connected to each other and to motor and sensory systems before birth and are operational in the human newborn. In addition, insular cortex, which is a key site for the convergence of exteroceptive and interoceptive signals, is likely to be involved in the early emergence of the self–world distinction and could encode an initial distinction between self and non-self based on interoception and self-touch [78,79]. Insular is the first cortical area to appear in the fetus and is considered to be both phylogenetically old and early maturing [80,81]. In figure 6, we highlight these structures, and particularly the SC and insular, as being potentially involved in a minimal, anoetic sense of self, and analogize them with layer 0 of our robot cognitive architecture (the correlation-learning layer). This proposal aligns with Rochat [5] and Rochat & Striano's [31] account of the emergence of a VSWD during early infancy that builds on a simpler non-visual distinction based on self-touch already present at birth. We next explore this neurodevelopmental proposal in more detail.

As shown in figure 6, the SC receives direct input from the retina alongside indirect input from auditory and somatosensory systems and a wide range of subcortical and cortical structures. SC is both a sensory and motor structure and is often studied in relation to orienting to salient stimuli and the control of spatial attention [82]. Indeed, the superficial layers of the SC are

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

well-known as encoding a head-centred representation of the world that aligns visual, auditory and tactile inputs [83,84], while the deeper layers act to generate eye and head movements to fixate or track objects of interest [76]. The SC is also known to provide corollary discharge signals for eye movements that help to stabilize visual perception [85]. However, while less studied, the SC and equivalent midbrain structures in non-mammalian vertebrates (the tectum) also play a wider role in movement generation (i.e. beyond orienting), including in reaching [86] and in whole body movement [87,88]. Interestingly, in primates, the deeper layers of the SC have been found to include cell populations that encode arm movements in gaze-related coordinates [89]. Other populations in SC, and the underlying PAG, show neural activity that correlates with shoulder, arm and trunk electromyography recordings during reaching [90], i.e. putative proprioceptive signals. Importantly, the SC does much more than simply provide a topographic map of the visual world. For instance, it is also involved in determining the salience of any objects in view with regard to future action [77,82]; indeed, in young infants, it is the primary site for this [91]. SC intrinsic circuitry has been modelled as implementing forms of correlation-based learning (e.g. [92,93]). Thus, in principle, SC seems a plausible site in the neonate brain for the acquisition of a VSWD, similar to that generated by layer 0 of our model, that could provide the developing infant with the capacity to identify areas of her visual world as belonging to self or world.

Cortical systems involved in visual processing mature rapidly during the first few month as they come to dominate the subcortical visual system [48,91]. There is evidence of function in visual cortical areas at birth [81], but the wider system is known to be incomplete and weakly integrated [91,94]. The brain expands in size during the first year, with cortical grey matter more than doubling, driven by the proliferation of intracortical synapses, myelination and increasing functional connectivity [94–96]. During this time, different cortical systems mature progressively, steered by experience [94,97] and beginning with sensorimotor and sensory areas, with a more gradual emergence of networks that process signals generated by these primary domains [96,98,99]. This aligns with theoretical proposals that cortical functional organization can be described by a hierarchy of predictive models [100].

From around three to four months of age, the infant brain gains the capacity to anticipate the future over short time frames, for instance, making prospective saccades towards the likely future position of a moving object or anticipating where an object might reappear that has become occluded (object permanence) [101,102]. As previously noted, predictively anticipating the sensory consequences of one's own movement is a key aspect of human agency [26,45,103]. Of the various cortical systems likely to be involved in a predictive VSWD the *inferior parietal lobe* (IPL) may be especially critical. In human adults, this area has been found to (i) encode the goal of motor actions [104], (ii) have topographic maps of the visual field and of the body [105,106], (iii) enable an allocentric body model that can be flexibly applied in representing the self or in reasoning about the actions of others [107], (iv) encode self-related body information [108], and (v) be active in studies that measure sense of agency [109,110]. As shown in figure 6, the IPL receives movement-related input from areas such as premotor cortex [104], information about the visual scene via the dorsal visual stream [111] and information about visually detected body parts from the extrastriate body area (EBA) [112]. It also reciprocally connects with the insular cortex [113], which, as we have already noted, may be one of the first cortical structures to be involved in the experience of self. With regard to its early development, there is evidence of functional activity related to action recognition in associative regions of cortex, including in areas that will become the IPL, emerging from around four to seven months [114,115].

Interestingly, the IPL has strong two-way connections with SC. Direct projections from IPL to SC terminate in the intermediate and deep layers [116,117], while IPL receives an indirect, topographically organized projection from SC superficial layers via the lateral pulvinar [116]. Input signals arriving in the lateral intraparietal area, which lies in the intraparietal sulcus (adjacent to the IPL), are not thought to be solely oculomotor but to be concerned with visual processing more generally [116]. Signals from SC may also be relayed to IPL via pulvinar projections to extrastriatal areas including the EBA [118]. It therefore seems plausible that the construction of a predictive VSWD, resembling that in layer 1 of our model, could involve the IPL and that this could build on subcortical representations of the visual field. This is consistent with imaging studies of the rubber hand illusion [119] that have implicated the SC as part of a broader system involved in bodily self-awareness in human adults.

A further important structure is the (hindbrain) cerebellum, which interacts with premotor cortex [120] and the IPL [116] via dysnaptic pathways involving the pontine nuclei and the thalamus. Its role in predicting sensory signals, including those related to self-movement, means that it is likely to be integral to the acquisition of a predictive VSWD [110]. The cerebellum also interacts reciprocally with the SC, again via the pontine nuclei [82,121]. The cerebellum has been modelled as an adaptive filter performing decorrelation [122], as being involved in calibrating collicular maps [123] and in sensory noise cancellation based on efference copy signals [124] (itself a form of self–world distinction). Cerebellar and pontine development is protracted but spikes during the early postnatal months. Although it is unclear to what extent the cerebellum could be involved in self–world discrimination in the earliest phase of infancy, its rapid post-birth development and interactions with key midbrain and cortical systems suggest that it could play an increasing, and potentially integrative, role as development proceeds.

(d) Scaffolded development in layered architectures

Our model implements two forms of staged learning both of which can be regarded as forms of scaffolding [43].

At level 0, during training of the correlation-based layer, we first focus on the *self* pathway and suppress training of the *world* pathway. We do this to ensure that the model learns to use proprioceptive signals to predict visual self-movement. Developmentally, this could be considered to be a form of attention (overt or covert), which prioritizes some forms of signal over others. Interestingly, the visual system of neonates is known to have very limited depth of field such that objects beyond 30 cm are poorly discriminated [48]. In learning terms, this could have the benefit of allowing the infant to focus on proximal stimuli such as her

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

own hand movements (also human faces), without being distracted by more distal stimuli. Infants are also known to spend time visually fixating their own hands [5], an activity that has been noted in newborns [125] and peaks before the onset of targeted reaching [36], thus 'self-structuring' [126] the visual signals provided to the developing brain.

Layer 1 scaffolding involves delaying training of the predicted layer until layer 0 training is completed and fixing the weights of the lower layer. This illustrates a broader principle that may apply elsewhere in human development, which is the capacity of representations acquired early on to scaffold learning in later maturing networks. In Wilson & Prescott [43], we discuss this in relation to the examples of infant face processing and mammalian thermoregulation. For example, there is good evidence of an early-developing subcortical face detection system, in the SC, that scaffolds the emergence of cortically based face recognition capacity [5,91,97]. We suggest that a similar form of scaffolding could be taking place in the emergence of the self–world distinction, again with the SC scaffolding learning in cortical structures such as IPL. Scaffolding is also likely to be important between cortical subsystems that form predictive hierarchies.

In the mammalian brain, there is significant postnatal development in the SC, particularly with respect to multi-sensory integration, regulated, at least in part, by top-down innervation from maturing cortical systems [127]. As noted above, representations in SC may also be modified and refined by developing SC-cerebellar circuitry. While we have focused in this discussion on scaffolding of the cortex by the sub-cortex, in Wilson & Prescott [43] we also discuss scaffolding of subcortical systems, such as the SC, by the cortex, but over much longer, evolutionary timescales. In evolving from the pre-mammalian tectum, we can expect that the SC will have changed under evolutionary pressure to better support cortical function, for instance, in constructing representations of the sensory world that are well-suited to act in consort with more predictively oriented cortical modules. In our architecture, the third term in equation (2.5), which enforces a form of temporal continuity in the layer 0 latent space (see analyses in the supplementary material), could be a model of this kind of evolution-imposed constraint.

In our analysis of model variants, we showed that our layered architecture approach is up to 98% more efficient than a prevailing machine learning benchmark (end-to-end BPTT) while still effectively segmenting the visual world and predicting future self-states. There are substantial efficiency gains even with the continuous training variant (see the supplementary material). Brain development is known to be energetically costly, with the brain in childhood requiring a significantly larger share of energy resource than in adults [128]. Indeed, the metabolic cost of brain development is considered to be an explanatory factor in the slow and protracted nature of childhood growth in humans compared to other mammals [128]. As training of AI models consumes an increasing share of the world's resources, we may be able to learn from the layered architecture of the human brain how to build and train very large network models more efficiently.

(e) Beyond the anoetic self

We envisage extending our synthetic model to include further components and tiers with the goal of capturing wider aspects of the anoetic self and, beyond that, of the noetic and autonoetic self. Here, we briefly consider some paths forward.

(i) Visually guided sensorimotor behaviour

As previously noted, hand and arm movements in early infancy are largely exploratory though they can be accompanied by reflexive grasping when an object touches the palm of the hand [34,37]. Intentional movements towards visual targets emerge from around age three months [35] and goal-oriented reach and grasp behaviour from around five to seven months [35,36]. This development sequence has already inspired a number of developmental robotics approaches (e.g. [129,130]). We foresee extending the current model along the following lines.

In order to predict future self-states, layer 1 of our architecture acquires a model that is functionally similar to inverse kinematics; however, to construct this, we need targeted movement. We currently use an engineered PID controller to generate babbling movements that bring body parts into the visual field. A more biomimetic approach, based on studies of spontaneous movement in the fetus [34], would involve generating pre-reaching movements using combinations of muscle/motor synergies similar to those generated in the brainstem and spinal cord [131,132]. Such synergy patterns could themselves form the target for layer 1 training in place of the direct proprioceptive states used in the current model (see [133] for evidence that cortical systems select muscle synergies). Note that what we are proposing here can again be viewed as a form of scaffolding in which subcortical systems generate movement patterns and representations that are useful targets for cortical learning.

A further extension is to use our architecture to generate reaching behaviours towards visual targets as seen in three-month-old infants. To do this, we need to map reaching targets in visual coordinates to the (x, y, z) world coordinates used in our layer 1 model. This model can then be used to generate movement trajectories to target locations, for instance, to reach towards visually salient objects. It should be possible to acquire the required mapping either during motor babbling or by driving the layer 1 model with a distribution of sample actions and using the imagined visual consequences to generate end positions in head-centred visual space. This problem is relatively straightforward in our current robot model as there are no separate degrees of freedom for the head, and it would become trivial if we add the depth signal from the robot RGB-D camera. For a robot with a head/neck morphology more similar to humans, the spatial locations of target stimuli will need to be updated with respect to head movements (e.g. [134]). The extension to goal-directed reaching, as observed in five- to seven-month-old infants, will require the integration of sensorimotor control with motivation as discussed next.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

(ii) Goal-directed behaviour

A further step will be to integrate evaluative and motivational systems into our architecture to build towards a more complete model of sense of agency and of goal-directed behaviour. In Prescott & Jimenez-Rodriguez [75], we analysed the mammalian brain architecture for motivation from a layered architecture perspective. This review identified several important subcortical substrates for motivated behaviour, including the medial reticular formation, the PAG, the hypothalamus and the basal ganglia. As noted above, these structures were also identified by Panksepp [11] as being implicated in the construction of a primal (anoetic) self equipped with 'a primitive sort of intentionality (an automatized action readiness) and primitive forms of psychic coherence (global affective states of the brain)' [11, p. 566]. Insular cortex is also considered to be a key forebrain structure for motivation, particularly in relation to homeostasis [78]. These structures likely contribute to the anoetic sense of agency, for instance, in relation to the experience of action authorship. Other cortical areas related to motivation are found in the medial and frontal cortical regions. These more slowly developing networks are usually associated with cognitive aspects of motivation, including representation of value, planning and monitoring, and are likely to be more involved in reflexive judgements about agency (see [75] and noetic self below). We expect to extend our layered architecture in the direction of goal-directed behaviour by building on some of our earlier neurorobotic models of relevant brain structures, including reticular formation [135], basal ganglia [136,137] and hypothalamus [138].

(iii) The self–other distinction

We have focused here on the self–world distinction which is likely to be a pre-requisite for the self–other distinction [3–5], where other is understood as a social other. Theories of the development of a self–other distinction typically begin with the hypothesis that the human brain uses self-representations as a model for understanding the actions and intentions of others [56,139–141]. Indeed, some of the brain structures in figure 6, such as premotor cortex and the IPL, are often found to be active in experimental settings that test for this kind of self–other mapping [140]. The possibility of using a robot self-model to distinguish self from a human or robot other, or to reason about the actions of a social other, has already been investigated in a number of studies [22,56,65,141]. The current model will lend itself naturally to adding a further decomposition of the world that distinguishes socially important cues such as facial expressions and body movements as a means to inferring intentions. We have reported some prior work in this direction [142–144].

(iv) The noetic self

A layered systems view of self, which is also consistent with predictive processing approaches [100] and with the hierarchical nature of human cortical development [99], would conceive of the progression towards the noetic self as involving the addition of new processing layers. Research by Tani & White [53] (reviewed above) suggests one possibility, which is the emergence of a noetic self through processes that monitor the anoetic self and respond to errors. For instance, when expectations about the reafferent sensory signals generated by self-movement are violated (see also [26]). Candidate structures in human cortex include the anterior cingulate, which has been proposed as playing a role in cognitive control of more automatic processes [145] as part of a wider 'salience' network [146]. As additional processing layers are added, representations of self can be expected to be increasingly multi-modal, or amodal, to encode expectations (priors) over longer time spans and to be the subject of self-referential thought. In the human brain, this conceptual representation of self likely involves the *default mode network* (DMN) [147] as discussed next.

(v) The autonoetic self

The ability to reflect on the self as existing in time, the *autonoetic* self, is generally held to emerge more gradually in development and to follow other childhood milestones, such as language acquisition and autobiographical memory [13,14,148]. From around age two, children show early signs of an awareness of a personal past and future, for instance, by reporting on past events and through appropriate use of the past tense of activity verbs. However, the capacity to fuse together autobiographical memories and self-concept using narrative-like structures is typically not in evidence until the pre-school years [14,148]. In the brain, this likely involves higher-order networks in the superior parietal lobe, the posterior cingulate cortex and the hippocampus. These structures form part of the DMN which is active during self-reflective thinking and mental time travel [147,149]. We expect to investigate these ideas by integrating our work on the self–world distinction with research on event memory systems where we have used cognitive robotics approaches to test models, inspired by the hippocampal system, that reconstruct the past and anticipate the future in embodied robotic settings [14,144].

4. Conclusion

We have argued that the human self–world distinction is acquired progressively and begins in an anoetic fashion (i.e. without self-awareness) with systems that learn to discriminate sensory signals emanating from (or generated by) the self, from those arising from the non-self. This distinction can guide attention, for example, towards the body for self-exploration, or away

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

from it, to better understand the world. A basic distinction of this form can also support learning of more sophistical forms of self–world distinction, for instance, those that predict how the future will evolve. Here, we have demonstrated a layered control architecture for a simulated humanoid robot that implements these two forms of self–world distinction—one correlational, the other predictive—and have shown scaffolded (faster and better) learning of the predictive model if the correlation layer is already in place and provided as input.

We have framed this modelling approach in the context of human brain development and infant behaviour. Specifically, we have argued that early infancy might involve the acquisition of a non-predictive VSWD, analagous to our layer 0 network, largely in subcortical areas of the brain, and partly through the infant's observation of her own motor babbling. During the first year of life, and as the brain further develops, this could lead to the acquisition of a second form of VSWD, cortically based, predictive in nature and analogous to layer 1 in our model. This two-layered architecture, still essentially anoetic, demonstrates an important aspect of agency, specifically the ability to predict the future consequences of one's own actions. The analogy with our robotic model follows a cognitive neurorobotics methodology and is focused at the systems level rather than in terms of specific mechanisms. Nevertheless, we hope that this decomposition into pathways and layers, alongside the distinction between correlational and predictive learning systems, could prove helpful in understanding the emergence of the human self–world distinction, particularly with respect to the role of early-developing self subsystems in scaffolding later ones.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Our paper uses computational models. Code is accessible via a public Github repository at: http://github.com/LucaManneschi/Robot_Self_Sim.git.

Supplementary material is available online [150].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. L.M.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; A.J-R.: conceptualization, formal analysis, investigation, methodology, software, validation, writing—review and editing; D.C.: conceptualization, investigation, methodology, software, visualization, writing—original draft; T.J.P.: conceptualization, funding acquisition, investigation, methodology, project administration, supervision, visualization, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. T.J.P. is a director and shareholder of Consequential Robotics Ltd. D.C. is a director and shareholder of Kinematic Trees Ltd. These companies are not expected to benefit from publication of this article. L.M. and A.J-R. declare that they have no competing interests.

Funding. A.J-R. was funded by Innovate UK (grant no. 10089807) under the UK's funding guarantee scheme for the EU Horizon Europe project PRIMI (grant agreement no. 101120727). T.J.P. was supported by Innovate UK (grant no. 10039052) under the UK's funding guarantee scheme for the EIC Pathfinder project CAVAA (grant agreement no. 101071178).

References

1. Jeannerod M, Pacherie E. 2004 Agency, simulation and self-identification. *Mind Lang.* **19**, 113–146. (doi:10.1111/j.1468-0017.2004.00251.x)
2. Gallagher S. 2000 Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* **4**, 14–21. (doi:10.1016/s1364-6613(99)01417-5)
3. Piaget J. 1952 *The origins of intelligence in the child*. New York, NY: International Universities Press.
4. Selman RL. 1980 *The growth of interpersonal understanding: developmental and clinical analyses*. New York, NY: Academic Press.
5. Rochat P. 2001 *The infant's world*. Cambridge, MA: Harvard University Press.
6. Simantov T, Lombardo M, Baron-Cohen S, Uzefovsky F. 2021 Self-other distinction. In *The neural basis of mentalizing* (eds M Gilead, KN Ochsner), pp. 85–106. Cham, Switzerland: Springer International Publishing.
7. Braitenberg V. 1986 Vehicles: experiments in synthetic psychology. Cambridge, MA: MIT Press.
8. Prescott TJ, Camilleri D. 2019 The synthetic psychology of the self. In *Cognitive architectures* (eds MIA Ferreira, JS Sequeira, R Ventura), pp. 85–104. Cham, Switzerland: Springer.
9. Verschure PFMJ, Prescott TJ. 2018 A living machines approach to the sciences of mind and brain. In *The handbook of living machines: research in biomimetic and biohybrid systems* (eds TJ Prescott, N Lepora, PFMJ Verschure), pp. 15–25. Oxford, UK: Oxford University Press.
10. Stout GF. 1896 *Analytic psychology*. London, UK: Swan Sonnenschein & Co.
11. Panksepp J. 1998 The periconscious substrates of consciousness: affective states and the evolutionary origins of the self. *J. Conscious. Stud.* **5**, 566–582. https://www.ingentaconnect.com/contentone/imp/jcs/1998/00000005/f0020005/895.
12. Fabbro F, Aglioti SM, Bergamasco M, Clarici A, Panksepp J. 2015 Evolutionary aspects of self- and world consciousness in vertebrates. *Front. Hum. Neurosci.* **9**, 157. (doi:10.3389/fnhum.2015.00157)
13. Tulving E. 1985 Memory and consciousness. *Can. J. Psychol.* **26**, 1–12. (doi:10.1037/h0080017)
14. Prescott TJ, Dominey PF. 2024 Synthesizing the temporal self: robotic models of episodic and autobiographical memory. *Phil. Trans. R. Soc. B* **379**, 20230415. (doi:10.1098/rstb.2023.0415)
15. Feinberg TE. 2001 *Altered egos: how the brain creates the self*. Oxford, UK: Oxford University Press.
16. Berkovich-Ohana A, Dor-Ziderman Y, Glicksohn J, Goldstein A. 2013 Alterations in the sense of time, space, and body in the mindfulness-trained brain: a neurophenomenologically-guided MEG study. *Front. Psychol.* **4**, 912. (doi:10.3389/fpsyg.2013.00912)
17. Deane G. 2020 Dissolving the self. *Philos. Mind Sci.* **1**, 1–27. (doi:10.33735/phimisci.2020.i.39)
18. Letheby C, Gerrans P. 2017 Self unbound: ego dissolution in psychedelic experience. *Neurosci. Conscious.* **2017**, nix016. (doi:10.1093/nc/nix016)
19. Botvinick M, Cohen J. 1998 Rubber hands 'feel' touch that eyes see. *Nature* **391**, 756–756. (doi:10.1038/35784)

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

20. Slater M, Spanlang B, Sanchez-Vives MV, Blanke O. 2010 First person experience of body transfer in virtual reality. *PLoS ONE* **5**, e10564. (doi:10.1371/journal.pone.0010564)
21. Martinez-Hernandez U, Boorman LW, Prescott TJ. 2017 Multisensory wearable interface for immersion and telepresence in robotics. *IEEE Sens. J.* **17**, 2534–2541. (doi:10.1109/jsen.2017.2669038)
22. Prescott TJ, Vogeley K, Wykowska A. 2024 Understanding the sense of self through robotics. *Sci. Robot.* **9**, eadn2733. (doi:10.1126/scirobotics.adn2733)
23. Moore K, Persaud T, Torchia M. 2015 *The developing human: clinically oriented embryology*. Amsterdam, The Netherlands: Elsevier.
24. Seth AK, Tsakiris M. 2018 Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* **22**, 969–981. (doi:10.1016/j.tics.2018.08.008)
25. Bahrick LE, Watson JS. 1985 Detection of intermodal proprioceptive–visual contingency as a potential basis of self-perception in infancy. *Dev. Psychol.* **21**, 963–973. (doi:10.1037/0012-1649.21.6.963)
26. Hohwy J. 2007 The sense of self in the phenomenology of agency and perception. *Psyche* **13**, 1–20. https://journalpsyche.org/files/0xab11.pdf.
27. Delafield-Butt J, Ciaunica A. 2024 Sensorimotor foundations of self-consciousness in utero. *Curr. Opin. Behav. Sci.* **59**, 101428. (doi:10.1016/j.cobeha.2024.101428)
28. Rochat P. 2019 Self-unity as ground zero of learning and development. *Front. Psychol.* **10**, 414. (doi:10.3389/fpsyg.2019.00414)
29. Rochat P, Hespos SJ. 1997 Differential rooting response by neonates: evidence for an early sense of self. *Early Dev. Parent.* **6**, 105–112. (doi:10.1002/(SICI)1099-0917(199709/12)6:3/4<105::AID-EDP150>3.0.CO;2-U)
30. von Holst E, Mittelstaedt H. 1950 Das reafferenzprinzip. *Naturwissenschaften* **37**, 464–476. (doi:10.1007/BF00622503)
31. Rochat P, Striano T. 2000 Perceived self in infancy. *Infant Behav. Dev.* **23**, 513–530. (doi:10.1016/S0163-6383(01)00055-8)
32. von Hofsten C. 1984 Developmental changes in the organization of prereaching movements. *Dev. Psychol.* **20**, 378–388. (doi:10.1037/0012-1649.20.3.378)
33. Wallace PS, Whishaw IQ. 2003 Independent digit movements and precision grip patterns in 1–5-month-old human infants: hand-babbling, including vacuous then self-directed hand and digit movements, precedes targeted reaching. *Neuropsychologia* **41**, 1912–1918. (doi:10.1016/S0028-3932(03)00128-3)
34. Fagard J, Esseily R, Jacquey L, O'Regan K, Somogyi E. 2018 Fetal origin of sensorimotor behavior. *Front. Neurorobot.* **12**, 23. (doi:10.3389/fnbot.2018.00023)
35. Heineman KR, Middelburg KJ, Hadders-Algra M. 2010 Development of adaptive motor behaviour in typically developing infants. *Acta Paediatr.* **99**, 618–624. (doi:10.1111/j.1651-2227.2009.01652.x)
36. White BL, Castle P, Held R. 1964 Observations on the development of visually-directed reaching. *Child Dev.* **35**, 349–364. (doi:10.2307/1126701)
37. Hadders-Algra M. 2018 Early human motor development: from variation to the ability to vary and adapt. *Neurosci. Biobehav. Rev.* **90**, 411–427. (doi:10.1016/j.neubiorev.2018.05.009)
38. Prescott TJ. 2015 Me in the machine. *New Sci.* **225**, 36–39. (doi:10.1016/S0262-4079(15)60554-1)
39. Damasio AR. 2010 *Self comes to mind: constructing the conscious brain*. London, UK: Vintage Books.
40. Panksepp J. 1998 *Affective neuroscience: the foundations of human and animal emotions*. Oxford: UK Oxford University Press.
41. 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. (doi:10.1016/j.neuroimage.2005.12.002)
42. Prescott TJ, Redgrave P, Gurney K. 1999 Layered control architectures in robots and vertebrates. *Adapt. Behav.* **7**, 99–127. (doi:10.1177/105971239900700105)
43. Wilson SP, Prescott TJ. 2022 Scaffolding layered control architectures through constraint closure: insights into brain evolution and development. *Phil. Trans. R. Soc. B* **377**, 20200519. (doi:10.1098/rstb.2020.0519)
44. Zelazo PD, Sommerville JA. 2001 Levels of consciousness of the self in time. In *The self in time: developmental perspectives* (eds C Moore, K Lemmon), pp. 229–252. Mahwah, NJ: Lawrence Erlbaum Associates.
45. David N, Newen A, Vogeley K. 2008 The 'sense of agency' and its underlying cognitive and neural mechanisms. *Conscious. Cogn.* **17**, 523–534. (doi:10.1016/j.concog.2008.03.004)
46. Bickhard MH. 2013 Scaffolding and self-scaffolding: central aspects of development. In *Children's development within social context, volume 1: metatheory and theory; volume 2: research and methodology* (eds LT Winegar, J Valsiner) pp. 33–52. Hove, UK: Psychology Press.
47. Tao S *et al.* 2025 ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI. In *Proc. Robotics: Science and Systems XXI*. Los Angeles, CA, USA. (doi:10.15607/RSS.2025.XXI.021)
48. Braddick O, Atkinson J. 2011 Development of human visual function. *Vis. Res.* **51**, 1588–1609. (doi:10.1016/j.visres.2011.02.018)
49. Milletari F, Navab N, Ahmadi SA. 2016 V-Net: fully convolutional neural networks for volumetric medical image segmentation. In *Proc. 2016 4th Int. Conf. 3D Vision (3DV)*, pp. 565–571. New York, NY: IEEE.
50. Kidger P, Foster J, Li X, Lyons TJ. 2021 Neural SDEs as infinite-dimensional GANs. In *Proc. Int. Conf. Machine Learning*, pp. 5453–5463. Cambridge, MA: PMLR.
51. Manneschi L *et al.* 2024 Noise-aware training of neuromorphic dynamic device networks. *Nat. Commun.* **16**, 9192. (doi:10.1038/s41467-025-64232-1)
52. Prescott TJ, Lepora N, Verschure PFMJ. 2018 *The handbook of living machines: research in biomimetic and biohybrid systems*. Oxford, UK: Oxford University Press.
53. Tani J, White J. 2022 Cognitive neurorobotics and self in the shared world, a focused review of ongoing research. *Adapt. Behav.* **30**, 81–100. (doi:10.1177/1059712320962158)
54. Prescott TJ, Wilson SP. 2023 Understanding brain functional architecture through robotics. *Sci. Robot.* **8**, eadg6014. (doi:10.1126/scirobotics.adg6014)
55. Cangelosi A, Schlesinger M, Smith LB. 2015 *Developmental robotics: from babies to robots*. Cambridge, MA: MIT Press.
56. Moller TJ, Georgie YK, Schillaci G, Voss M, Hafner VV, Kaltwasser L. 2021 Computational models of the 'active self' and its disturbances in schizophrenia. *Conscious. Cogn.* **93**, 103155. (doi:10.1016/j.concog.2021.103155)
57. Prescott TJ. 2021 Robot self. In *Encyclopedia of robotics* (eds MH Ang, O Khatib, B Siciliano), pp. 1–9. Berlin, Germany: Springer.
58. Tani J. 1998 An interpretation of the 'self' from the dynamical systems perspective: a constructivist approach. *J. Conscious. Stud.* **5**, 516–542. https://www.ingentaconnect.com/contentone/imp/jcs/1998/00000005/f0020005/880.
59. Michel P, Gold K, Scassellati B. 2004 Motion-based robotic self-recognition. In *Proc. 2004 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 2763–2768. New York, NY: IEEE.
60. Saegusa R, Metta G, Sandini G. 2010 Own body perception based on visuomotor correlation. In *Proc. 2010 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 1044–1051. New York, NY: IEEE.
61. Demirel B, Moulin-Frier C, Arsiwalla XD, Verschure PFMJ, Sánchez-Fibla M. 2021 Distinguishing self, other, and autonomy from visual feedback: a combined correlation and acceleration transfer analysis. *Front. Hum. Neurosci.* **15**, 560657. (doi:10.3389/fnhum.2021.560657)
62. Bechtle S, Schillaci G, Hafner VV. 2016 On the sense of agency and of object permanence in robots. In *Proc. Joint IEEE Int. Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 166–171. New York, NY: IEEE.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

63. Lanillos P, Dean-Leon E, Cheng G. 2017 Yielding self-perception in robots through sensorimotor contingencies. *IEEE Trans. Cogn. Dev. Syst.* **9**, 100–112. (doi:10.1109/TCDS.2016.2627820)
64. Lanillos P, Cheng G. 2018 Adaptive robot body learning and estimation through predictive coding. In *Proc. 2018 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 4083–4090. New York, NY: IEEE.
65. Lanillos P, Pages J, Cheng G. 2020 Robot self/other distinction: active inference meets neural networks learning in a mirror. In *Proc. 24th Eur. Conf. Artificial Intelligence*, pp. 2410–2416. Amsterdam, The Netherlands: IOS Press.
66. Meo C, Lanillos P. 2021 Multimodal VAE active inference controller. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2693–2699. (doi:10.1109/IROS51168.2021.9636394)
67. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. 2018 Neural ordinary differential equations. In *Proc. Int. Conf. Neural Information Processing Systems*, pp. 6572–6583. Red Hook, NY: Curran Associates Inc.
68. Rubanova Y, Chen RTQ, Duvenaud D. 2019 Latent ODEs for irregularly-sampled time series. In *Proc. 33rd Int. Conf. Neural Information Processing Systems*, pp. 5320–5353. Red Hook, NY: Curran Associates Inc.
69. Yıldız Ç, Heinonen M, Lähdesmäki H. 2019 ODE2VAE: deep generative second order ODEs with Bayesian neural networks. In *Proc. 32nd Int. Conf. Neural Information Processing Systems*, pp. 13412–13421. Red Hook, NY: Curran Associates Inc.
70. Toth P, Rezende DJ, Jaegle A, Racaniere S, Botev A, Higgins I. 2020 Hamiltonian generative networks. In *Proc. 8th Int. Conf. Learning Representations (ICLR)*.
71. Park S, Kim K, Lee J, Choo J, Lee J, Kim S, Choi E. 2021 Vid-ODE: continuous-time video generation with neural ordinary differential equation. In *Proc. 35th AAAI Conf. Artificial Intelligence*, pp. 2412–2422. Washington, DC: AAAI Press. (doi:10.1609/aaai.v35i3.16342)
72. Kanaa D, Voleti V, Ebrahimi Kahou S, Pal C. 2021 Simple video generation using neural ODEs. *arXiv*. (doi:10.48550/arXiv.2109.03292)
73. Hafner D, Lillicrap T, Ba J, Norouzi M. 2019 Dream to control: learning behaviors by latent imagination. *arXiv*. (doi:10.48550/arXiv.1912.01603)
74. Hafner D, Pasukonis J, Ba J, Lillicrap T. 2023 Mastering diverse domains through world models. *Nature* **640**, 647–653. (doi:10.1038/s41586-025-08744-2)
75. Prescott TJ, Jimenez-Rodriguez A. 2025 Understanding the layered brain architecture for motivation: dynamical systems, computational neuroscience, and robotic approaches. In *Psychology of learning and motivation: intelligence in a physical wo rld, volume 82* (eds KD Federmeier, JOS Goh), pp. 45–96. Cambridge, MA: Academic Press.
76. Gandhi N, Katnani H. 2011 Motor functions of the superior colliculus. *Annu. Rev. Neurosci* **34**, 205–231. (doi:10.1146/annurev-neuro-061010-113728)
77. Merker B. 2007 Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav. Brain Sci.* **30**, 63–81. (doi:10.1017/S0140525X07000891)
78. Craig AN. 2015 *How do you feel?: an interoceptive moment with your neurobiological self*. Princeton, NJ: Princeton University Press.
79. Tisserand A, Philippi N, Botzung A, Blanc F. 2023 Me, myself and my insula: an oasis in the forefront of self-consciousness. *Biology* **12**, 599. (doi:10.3390/biology12040599)
80. Meyer G. 2018 From the lateral edge to the center of the cortex: the development of the human insula. *Neuroforum* **24**, A151–A158. (doi:10.1515/nf-2018-A008)
81. Lemaître H *et al.* 2021 Rest functional brain maturation during the first year of life. *Cereb. Cortex* **31**, 1776–1785. (doi:10.1093/cercor/bhaa325)
82. Krauzlis RJ, Lovejoy LP, Zénon A. 2013 Superior colliculus and visual spatial attention. *Annu. Rev. Neurosci.* **36**, 165–182. (doi:10.1146/annurev-neuro-062012-170249)
83. Sparks DL, Nelson IS. 1987 Sensory and motor maps in the mammalian superior colliculus. *Trends Neurosci.* **10**, 312–317. (doi:10.1016/0166-2236(87)90085-3)
84. Prabhu NG, Himmelbach M. 2023 Activity in the human superior colliculus associated with reaching for tactile targets. *NeuroImage* **280**, 120322. (doi:10.1016/j.neuroimage.2023.120322)
85. Sommer MA, Wurtz RH. 2002 A pathway in primate brain for internal monitoring of movements. *Science* **296**, 1480–1482. (doi:10.1126/science.1069590)
86. Philipp R, Hoffmann K-P. 2014 Arm movements induced by electrical microstimulation in the superior colliculus of the macaque monkey. *J. Neurosci.* **34**, 3350–3363. (doi:10.1523/JNEUROSCI.0443-13.2014)
87. Dean P, Redgrave P, Westby GWM. 1989 Event or emergency? Two response systems in the mammalian superior colliculus. *Trends Neurosci.* **12**, 137–147. (doi:10.1016/0166-2236(89)90052-0)
88. Cooper B, McPeek RM. 2021 Role of the superior colliculus in guiding movements not made by the eyes. *Annu. Rev. Vis. Sci.* **7**, 279–300. (doi:10.1146/annurev-vision-012521-102314)
89. Stuphorn V, Bauswein E, Hoffmann KP. 2000 Neurons in the primate superior colliculus coding for arm movements in gaze-related coordinates. *J. Neurophysiol.* **83**, 1283–1299. (doi:10.1152/jn.2000.83.3.1283)
90. Werner W, Dannenberg S, Hoffmann KP. 1997 Arm-movement-related neurons in the primate superior colliculus and underlying reticular formation: comparison of neuronal activity with EMGs of muscles of the shoulder, arm and trunk during reaching. *Exp. Brain Res.* **115**, 191–205. (doi:10.1007/PL00005690)
91. Johnson MH. 1990 Cortical maturation and the development of visual attention in early infancy. *J. Cogn. Neurosci.* **2**, 81–95. (doi:10.1162/jocn.1990.2.2.81)
92. Cuppini C, Magosso E, Rowland B, Stein B, Ursino M. 2012 Hebbian mechanisms help explain development of multisensory integration in the superior colliculus: a neural network model. *Biol. Cybern.* **106**, 691–713. (doi:10.1007/s00422-012-0511-9)
93. Pitti A, Kuniyoshi Y, Quoy M, Gaussier P. 2013 Modeling the minimal newborn's intersubjective mind: the visuotopic-somatotopic alignment hypothesis in the superior colliculus. *PLoS ONE* **8**, e69474. (doi:10.1371/journal.pone.0069474)
94. Li M, Liu T, Xu X, Wen Q, Zhao Z, Dang X, Zhang Y, Wu D. 2022 Development of visual cortex in human neonates is selectively modified by postnatal experience. *eLife* **11**, e78733. (doi:10.7554/eLife.78733)
95. Yin W *et al.* 2025 Charting brain functional development from birth to 6 years of age. *Nat. Hum. Behav.* **9**, 1246–1259. (doi:10.1038/s41562-025-02160-2)
96. Gilmore JH, Knickmeyer RC, Gao W. 2018 Imaging structural and functional brain development in early childhood. *Nat. Rev. Neurosci.* **19**, 123–137. (doi:10.1038/nrn.2018.1)
97. Johnson MH. 2011 Interactive specialization: a domain-general framework for human functional brain development? *Dev. Cogn. Neurosci.* **1**, 7–21. (doi:10.1016/j.dcn.2010.07.003)
98. Gao W, Alcauter S, Elton A, Hernandez-Castillo CR, Smith JK, Ramirez J, Lin W. 2015 Functional network development during the first year: relative sequence and socioeconomic correlations. *Cereb. Cortex* **25**, 2919–2928. (doi:10.1093/cercor/bhu088)
99. Gao W. 2025 A hierarchical model of early brain functional network development. *Trends Cogn. Sci.* **29**, 855–868. (doi:10.1016/j.tics.2025.04.001)
100. Bastos A, Usrey W, Adams R, Mangun G, Fries P, Friston K. 2012 Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711. (doi:10.1016/j.neuron.2012.10.038)
101. von Hofsten C, Vishton P, Spelke ES, Feng Q, Rosander K. 1998 Predictive action in infancy: tracking and reaching for moving objects. *Cognition* **67**, 255–285. (doi:10.1016/S0010-0277(98)00029-8)
102. von Hofsten C. 2004 An action perspective on motor development. *Trends Cogn. Sci.* **8**, 266–272. (doi:10.1016/j.tics.2004.04.002)

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

103. Jeannerod M. 2003 The mechanism of self-recognition in humans. *Behav. Brain Res.* **142**, 1–15. (doi:10.1016/S0166-4328(02)00384-4)
104. Bonini L, Rozzi S, Serventi FU, Simone L, Ferrari PF, Fogassi L. 2010 Ventral premotor and inferior parietal cortices make distinct contribution to action organization and intention understanding. *Cereb. Cortex* **20**, 1372–1385. (doi:10.1093/cercor/bhp200)
105. Snyder LH, Grieve KL, Brotchie P, Andersen RA. 1998 Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394**, 887–891. (doi:10.1038/29777)
106. Silver MA, Kastner S. 2009 Topographic maps in human frontal and parietal cortex. *Trends Cogn. Sci.* **13**, 488–495. (doi:10.1016/j.tics.2009.08.005)
107. Farrer C, Frith CD. 2002 Experiencing oneself vs. another person as being the cause of an action: the neural correlates of the experience of agency. *NeuroImage* **15**, 596–603. (doi:10.1006/nimg.2001.1009)
108. Hodzic A, Muckli L, Singer W, Stirn A. 2009 Cortical responses to self and others. *Hum. Brain Mapp.* **30**, 951–962. (doi:10.1002/hbm.20558)
109. Blakemore SJ, Frith C. 2003 Self-awareness and action. *Curr. Opin. Neurobiol.* **13**, 219–224. (doi:10.1016/S0959-4388(03)00043-6)
110. Blakemore SJ, Sirigu A. 2003 Action prediction in the cerebellum and in the parietal lobe. *Exp. Brain Res.* **153**, 239–245. (doi:10.1007/s00221-003-1597-z)
111. Kravitz DJ, Saleem KS, Baker CI, Mishkin M. 2011 A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.* **12**, 217–230. (doi:10.1038/nrn3008)
112. Zimmermann M, Mars RB, de Lange FP, Toni I, Verhagen L. 2018 Is the extrastriate body area part of the dorsal visuomotor stream? *Brain Struct. Funct.* **223**, 31–46. (doi:10.1007/s00429-017-1469-0)
113. Uddin LQ, Nomi JS, Hébert-Seropian B, Ghaziri J, Boucher O. 2017 Structure and function of the human insula. *J. Clin. Neurophysiol.* **34**. 300–306. (doi:10.1097/wnp.0000000000000377)
114. Grossmann T, Cross ES, Ticini LF, Daum MM. 2013 Action observation in the infant brain: the role of body form and motion. *Soc. Neurosci.* **8**, 22–30. (doi:10.1080/17470919.2012.696077)
115. Mento G, Duma GM, Valenza E, Farroni T. 2022 Face specific neural anticipatory activity in infants 4 and 9 months old. *Sci. Rep.* **12**, 12938. (doi:10.1038/s41598-022-17273-1)
116. Clower DM, West RA, Lynch JC, Strick PL. 2001 The inferior parietal lobule is the target of output from the superior colliculus, hippocampus, and cerebellum. *J. Neurosci.* **21**, 6283–6291. (doi:10.1523/JNEUROSCI.21-16-06283.2001)
117. Borra E, Gerbella M, Rozzi S, Tonelli S, Luppino G. 2014 Projections to the superior colliculus from inferior parietal, ventral premotor, and ventrolateral prefrontal areas involved in controlling goal-directed hand actions in the macaque. *Cereb. Cortex* **24**, 1054–1065. (doi:10.1093/cercor/bhs392)
118. Arcaro MJ, Pinsk MA, Kastner S. 2015 The anatomical and functional organization of the human visual pulvinar. *J. Neurosci.* **35**, 9848–9871. (doi:10.1523/JNEUROSCI.1575-14.2015)
119. Olivé I, Tempelmann C, Berthoz A, Heinze HJ. 2015 Increased functional connectivity between superior colliculus and brain regions implicated in bodily self-consciousness during the rubber hand illusion. *Hum. Brain Mapp.* **36**, 717–730. (doi:10.1002/hbm.22659)
120. Ramnani N. 2006 The primate cortico-cerebellar system: anatomy and function. *Nat. Rev. Neurosci.* **7**, 511–522. (doi:10.1038/nrn1953)
121. May PJ. 2006 The mammalian superior colliculus: laminar structure and connections. *Prog. Brain Res.* **151**, 321–378. (doi:10.1016/S0079-6123(05)51011-2)
122. Dean P, Porrill J. 2008 Adaptive-filter models of the cerebellum: computational analysis. *Cerebellum* **7**, 567–571. (doi:10.1007/s12311-008-0067-3)
123. Wilson ED, Anderson SR, Dean P, Porrill J. 2019 Sensorimotor maps can be dynamically calibrated using an adaptive-filter model of the cerebellum. *PLOS Comput. Biol.* **15**, e1007187. (doi:10.1371/journal.pcbi.1007187)
124. Anderson SR, Porrill J, Prescott TJ, Dean P. 2012 An internal model architecture for novelty detection: implications for cerebellar and collicular roles in sensory processing. *PLoS ONE* **7**, e44560. (doi:10.1371/journal.pone.0044560)
125. Rochat P. 1998 Self-perception and action in infancy. *Exp. Brain Res.* **123**, 102–109. (doi:10.1007/s002210050550)
126. Lungarella M, Sporns O. 2005 Information self-structuring: key principle for learning and development. In *Proc. 4th Int. Conf. Development and Learning*, pp. 25–30. New York, NY: IEEE.
127. Wallace M, Stein B. 1994 Cross-modal synthesis in the midbrain depends on input from cortex. *J. Neurophysiol.* **71**, 429–432. (doi:10.1152/jn.1994.71.1.429)
128. Kuzawa CW *et al.* 2014 Metabolic costs and evolutionary implications of human brain development. *Proc. Natl Acad. Sci. USA* **111**, 13 010–13 015. (doi:10.1073/pnas.1323099111)
129. Law J, Shaw P, Earland K, Sheldon M, Lee MH. 2014 A psychology based approach for longitudinal development in cognitive robotics. *Front. Neurorobot.* **8**, 1. (doi:10.3389/fnbot.2014.00001)
130. Juett J, Kuipers B. 2019 Learning and acting in peripersonal space: moving, reaching, and grasping. *Front. Neurorobot.* **13**, 4. (doi:10.3389/fnbot.2019.00004)
131. Bizzi E, Cheung V, d'Avella A, Saltiel P, Tresch M. 2008 Combining modules for movement. *Brain Res. Rev.* **57**, 125–133. (doi:10.1016/j.brainresrev.2007.08.004)
132. Saito H, Yokoyama H, Sasaki A, Kato T, Nakazawa K. 2022 Evidence for basic units of upper limb muscle synergies underlying a variety of complex human manipulations. *J. Neurophysiol.* **127**, 958–968. (doi:10.1152/jn.00499.2021)
133. Rana M, Yani MS, Asavasopon S, Fisher BE, Kutch JJ. 2015 Brain connectivity associated with muscle synergies in humans. *J. Neurosci.* **35**, 14 708–14 716. (doi:10.1523/JNEUROSCI.1971-15.2015)
134. Bayer M, Zimmermann E. 2024 Active head movements contribute to spatial updating across gaze shifts. *R. Soc. Open Sci.* **11**, 231545. (doi:10.1098/rsos.231545)
135. Humphries M, Gurney K, Prescott TJ. 2005 Is there an integrative center in the vertebrate brain-stem? A robotic evaluation of a model of the reticular formation viewed as an action selection device. *Adapt. Behav.* **13**, 97–113. (doi:10.1177/105971230501300203)
136. Prescott TJ, Gonzalez FM, Humphries MD, Gurney K, Redgrave P. 2006 A robot model of the basal ganglia: behaviour and intrinsic processing. *Neural Netw.* **19**, 31–61. (doi:10.1016/j.neunet.2005.06.049)
137. Prescott TJ, Montes González FM, Gurney K, Humphries MD, Redgrave P. 2024 Simulated dopamine modulation of a neurorobotic model of the basal ganglia. *Biomimetics* **9**, 139. (doi:10.3390/biomimetics9030139)
138. Jimenez-Rodriguez A, Prescott TJ. 2023 Motivational modulation of consummatory behaviour and learning in a robot model of spatial navigation. In *Biomimetic and biohybrid systems* (eds F Meder, A Hunt, L Margheri, A Mura, B Mazzolai), pp. 240–253. Cham, Switzerland: Springer Nature.
139. Meltzoff AN, Brooks R 'Like me' as a building block for understanding other minds: bodily acts, attention, and intention. In *Intentions and intentionality: foundations of social cognition* (eds BF Malle, LJ Moses, DA Baldwin), pp. 171–191. Cambridge, MA: MIT Press.
140. Decety J, Sommerville JA. 2003 Shared representations between self and other: a social cognitive neuroscience view. *Trends Cogn. Sci.* **7**, 527–533. (doi:10.1016/j.tics.2003.10.004)

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026

141. Smith LB, Breazeal C. 2007 The dynamic lift of developmental process. *Dev. Sci.* **10**, 61–68. (doi:10.1111/j.1467-7687.2007.00565.x)
142. Camilleri D, Prescott TJ. 2017 Action recognition with unsynchronised multi-sensory data. In *Proc. 7th Joint IEEE Int. Conf. Developmevt and Learning and Epigenetic Robotics (ICDL-EPIROB)*, pp. 53–59. New York, NY: IEEE.
143. Moulin-Frier C *et al.* 2018 DAC-h3: a proactive robot cognitive architecture to acquire and express knowledge about the world and the self. *IEEE Trans. Cogn. Dev. Syst.* **10**, 1005–1022. (doi:10.1109/TCDS.2017.2754143)
144. Prescott TJ, Camilleri D, Martinez-Hernandez U, Damianou A, Lawrence ND. 2019 Memory and mental time travel in humans and social robots. *Phil. Trans. R. Soc. B* **374**, 20180025. (doi:10.1098/rstb.2018.0025)
145. Shenhav A, Cohen JD, Botvinick MM. 2016 Dorsal anterior cingulate cortex and the value of control. *Nat. Neurosci.* **19**, 1286–1291. (doi:10.1038/nn.4384)
146. Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, Reiss AL, Greicius MD. 2007 Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* **27**, 2349–2356. (doi:10.1523/JNEUROSCI.5587-06.2007)
147. Davey CG, Pujol J, Harrison BJ. 2016 Mapping the self in the brain's default mode network. *NeuroImage* **132**, 390–397. (doi:10.1016/j.neuroimage.2016.02.022)
148. Fivush R. 2011 The development of autobiographical memory. *Annu. Rev. Psychol.* **62**, 559–582. (doi:10.1146/annurev.psych.121208.131702)
149. Menon V. 2023 20 years of the default mode network: a review and synthesis. *Neuron* **111**, 2469–2487. (doi:10.1016/j.neuron.2023.04.023)
150. Manneschi L, Jimenez-Rodriguez A, Camilleri D, Prescott TJ. 2026Supplementary material from: Scaffolded acquisition of a robot self–world distinction in a layered architecture model of the developing self. Figshare. (doi:10.6084/m9.figshare.c.8721458)

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0153/6176876/rstb.2025.0153.pdf
by guest
on 13 November 2026