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/247956/

Version: Published Version

Article:
Prescott, T.J., Wykowska, A., Garfinkel, S.N. et al. (2026) From being to identity: analysis and synthesis of the self. Philosophical Transactions of the Royal Society B: Biological Sciences, 381 (1962). 20250152. ISSN: 0962-8436

https://doi.org/10.1098/rstb.2025.0152

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

Introduction

Cite this article: Prescott TJ, Wykowska A, Garfinkel SN, Verschure PFMJ. 2026 From being to identity: analysis and synthesis of the self. *Phil. Trans. R. Soc. B* **381**: 20250152.
https://doi.org/10.1098/rstb.2025.0152

Received: 29 September 2026
Accepted: 29 September 2026

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

Subject Areas:
behaviour, cognition, neuroscience

Keywords:
sense of self, self-world distinction, agency, subjectivity, interoception, robotics, synthetic selves

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

From being to identity: analysis and synthesis of the self

Tony J. Prescott[1], Agnieszka Wykowska[2], Sarah N. Garfinkel[3] and Paul F. M. J. Verschure[4]

[1]School of Computer Science, The University of Sheffield, Sheffield, UK
[2]Center for Human Technologies, Istituto Italiano di Tecnologia, Genova, Italy
[3]Institute of Cognitive Neuroscience, University College London, London, UK
[4]CSIC-UMH Institute of Neuroscience and Department of Health Psychology, Universidad Miguel Hernández, Elche, Spain

TJP, 0000-0003-4927-5390; AW, 0000-0003-3323-7357; SNG, 0000-0002-5961-1012; PFMJV, 0000-0003-3643-9544

The experience of being, or having, a self—contained by our bodies and able to act in the world—comes naturally to us as humans, along with a feeling of being the same self from day to day and of seeing others as also being selves. But what is the self, and how is it constituted? In this theme issue, we explore these questions through contemporary *analytic* approaches, in philosophy, psychology and neuroscience, that seek to understand the human self through conceptual and empirical investigations. Alongside this, we also explore *synthetic* approaches that attempt to construct a sense of self, or some part thereof, in artefacts such as robots. Our overall aim is to deepen the understanding of self as a foundational aspect of the human condition and explore both theoretical and practical implications of the emerging science of the self. In this introduction, we review the scope and content of the theme issue and argue that these complementary research strands can be seen as converging towards a 'systems' view that sees the self as an emergent property of an embodied, dynamical and informational system that continuously distinguishes, regulates and models itself through processes that sustain and explain its own existence.

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

1. Introduction – what is the self?

While Descartes famously asserted that the self was simple and distinct from the body [1, vol. VII, p. 78], he also presented a second conception of self as a composite equated with the whole person [1, vol. VII, pp. 81–82] (see also [2]). This tension between the intuition of the self as an unanalysable essence and that of the self as an aggregate enmeshed with the body is reflected in hundreds of years of debate in science and philosophy about the human condition. For William James, there were two sides to the self—the subject 'I' and the object 'me' [3]. While philosophy and consciousness science have largely focused on the problem of subjectivity, psychology and, latterly, cognitive neuroscience have developed a science of understanding self as object. An important theme, initiated by Ulrich Neisser [4–6], is the decomposition of the self into ecological (embodied), interpersonal, temporally extended, conceptual and private aspects. A further line of research has focused on the narrative self [7–9]—the construction of a coherent and meaningful identity through storytelling. These different stances reflect not only the different focuses of neuroscience, philosophy and psychology as disciplines

© 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.

THE ROYAL SOCIETY PUBLISHING

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

but also that the self has these multiple aspects—as the primary organizer of sensation and generator of action, as the locus of phenomenological experience and as the central player in our personal life stories.

Recent results from developmental psychology, cognitive neuroscience and neurology, some reported in this issue [10–13], support the notion of dissociable self subsystems that combine to form the broader sense of self [4,14–18], while ongoing synthetic approaches, including the construction of physical models (robots), are making it possible to build and test theories of self in behaving systems [14,19–24]. By bringing together findings from philosophy, psychology, neuroscience, artificial intelligence and robotics, we hope to demonstrate the value of exploring the construction of the self from both analytic and synthetic perspectives. In this introduction, we will also propose that there is potential for convergence. Specifically, we will argue that a 'systems' view of self can reunify the 'I' and 'me' of the self-duality and of the self as both distinct from, yet enmeshed with, the body. This view sees the self as an emergent property of an embodied, dynamical and informational system that continuously maintains its organization, regulates its own internal state and sensorimotor engagement with the environment and progressively constructs models of itself at increasingly explicit levels. A second key theme is that the human self should also be seen as profoundly relational and extended (beyond the physical body and the here and now). That is, the perspectives of social others and of wider culture are internalized by the developing self to form part of its core architecture, particularly in relation to its normative and narrative aspects.

2. Analysis and synthesis of the self

While the self has been studied in multiple fields and under different guises, the past decade has led to a renewed interest in a science of the self, both fundamental and applied. This has come about both through the emergence of new conceptual frameworks (e.g. [18,21,25–29]) and novel methodologies for understanding the self, including (i) advances in systems neuroscience and whole brain imaging (e.g. [30–34]), (ii) technologies for transforming the self, including virtual reality, brain–machine interfaces, telepresence and teleoperation robotics (e.g. [10,31,35]) and (iii) renewed interest in the self-transforming capabilities of psychoactive drugs (e.g. [36–39]). A common theme of much of this work is to see the self as a process, or, more accurately as a cluster of interacting processes, from which the different aspects of self emerge.

Our strong belief is that progress in the biological sciences, and in the engineering of intelligent adaptive systems, can be accelerated through the convergence of analytic and synthetic approaches [40]. Indeed, a key paradigm across several contributions to this theme issue is the use of cognitive neurorobotics to explore self-construction [19,20,29]. In this work, theoretical principles are translated into computational frameworks and deployed in robots to observe emergent self-related signatures. This methodology can create a testable alignment between philosophical concepts such as ego and agency and functional models derived from psychology and neuroscience. Within this approach, an emerging view is that the brain develops body-grounded models of the self and the world as the basis for an initial self–world distinction [20,21,41]. As more complex models are constructed through development, the self forms a *layered* or hierarchical system derived from generative processes cast as predictive inference [20,42,43] or as prediction-driven control [29]. The functional roles of different subsystems can be more precisely characterized by this approach, as can the emergent consequences of their interactions.

Recent methodological approaches in cognitive science have also shown that robots can be used as sophisticated experimental apparatus to study some of the mechanisms underlying the human sense of self [21]. Being physically embodied, robots provide greater ecological validity than artificial stimuli presented on a screen. Moreover, in experiments with robots (see, for instance [10], this issue), participants can engage in a real-time exchange of social signals and in joint action in a highly controllable manner. This provides a powerful new means for studying mechanisms of social cognition that underlie the interpersonal aspects of self.

3. Evolving, developing and diverse selves

It can be argued that a multilayered self architecture is an evolutionary necessity [29,44]. For instance, biological agency can be framed as driven by physical survival first and subsequently by the transition to social navigation. Early control systems evolved to address the challenge of physical foraging, framed by Verschure [45, this issue] as solving the H4W problem: 'how, why, what, where, when'. This level is characterized by 'individual agency'—self-regulation for individual goal pursuit through feedback control—and already requires the distinction of self from non-self [19,46–48]. The evolution of complex metazoan life led to animal societies, and this emergence of social interactivity added a fifth question, 'who' (H5W), requiring qualitatively different control architectures [45]. Human psychology is particularly group-oriented and may have evolved through two distinct steps in social agency [49, this issue]: 'joint agency' (c. 2 Mya)—the self-reflective capacity to identify as a group member—and 'collective agency' (c. 130 kya)—the self-regulatory motivation to meet group standards.

This evolutionary trajectory of the sense of self may have interesting parallels in the emergence of self in the developing human mind.

Research summarized by Manneschi *et al.* [19, this issue] suggests the human sense of self is underpinned by a layered cognitive architecture from which different aspects of self emerge at different developmental stages. Core aspects of agency and ownership are available early, grounded in subcortical structures. For example, neonates can distinguish self-touch from external touch [50]. The early self–world distinction rapidly broadens as the infant develops. For instance, through observing its own actions

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

during motor babbling, the infant can learn dependencies between proprioception and vision to segregate 'me' from 'not me' in the visual domain. As Manneschi *et al.* demonstrate, using machine learning and robot simulation, this core system can then scaffold the development of anticipatory systems that forecast how self and world might evolve in the future.

Southgate [11, this issue] reviews evidence that self-awareness is anchored in infants' interoceptive signals (e.g. heartbeat, breathing, etc). Sensitivity to these signals is modulated by caregiver attention (e.g. direct gaze), which alerts the infant to their internal states. Initially, a 'skeleton' concept of self emerges—a representation of the self as an object that can perceive and be perceived—to be later enriched with identity, roles and preferences. By the end of the first year, infants understand gaze as referential. When a caregiver gazes at them, infants seek a 'target' and fill this conceptual placeholder with the 'belonging-to-the-body' feeling state triggered by the interaction.

Southgate also discusses mirror self-recognition (MSR) [51], which emerges by around two years of age, as marking a further progression in self-referential processing. The timing of milestones like MSR varies across cultures, potentially owing to differences in the degree of face-to-face non-verbal communication between infants and carers [52]. Self-perception continues to change through puberty (e.g. [53, this issue]) and beyond.

Around age 4–5 years, the capacity for mental time travel arises alongside the acquisition of autobiographical memory, allowing the child to fuse self-concept with a narrative personal history [54]. Carers help children organize experience into connected narratives (e.g. bedtime reminiscence), a form of social scaffolding, which leads to the construction of a personal timeline [11]. This self-narration is progressively consolidated and updated through social feedback to form the mature narrative self [55, this issue].

The study of neurodiversity can also provide valuable evidence of the constructed and decomposable nature of the self. For example, disorders, such as personal neglect and limb alienation, provide evidence that the embodied self depends on an internal representation of the body that can become dissociated from the physical body [21,56]. Damage to the hippocampal system can severely impair episodic memory and mental time travel, impacting the experience of the self as existing across time [57]. However, individuals with severe episodic memory loss (e.g. Alzheimer's or sudden-onset amnesia) often report a preserved sense of phenomenological continuity [58, this issue]. This suggests that self-continuity is not solely mnemonic but is sustained by interoception, embodiment and a stable life trajectory (see also [18,42]).

Other forms of neurodiversity also reveal differential processing of the self. For instance, Vogeley *et al.* [13, this issue] explore how autism may reflect an altered development of the interpersonal self, specifically through differences in multimodal coordination and interpersonal alignment. Major depressive disorder (MDD) involves persistent, self-critical processing and the formation of negative self-schemas (e.g. beliefs of worthlessness). Clinical trials using psilocybin, including a new study by Belinger *et al.* [39, this issue], have demonstrated that modulating the neural correlates of self-referential processing could alleviate symptoms of MDD by reducing ruminative self-focus and fostering a more positive self-concept.

4. Towards a convergent view of self

While the various contributions to this theme issue are often concerned with different aspects or notions of self, it is also possible to see convergence.

First, there is no support for a Cartesian 'substantialist' view of the self as a detached and unchanging essence, which, as Day and Zahavi note [38, this issue], is a straw man position that is not defended in contemporary philosophy (and see [2] for a discussion of Descartes' multi-faceted view of self). Instead, the self is understood as real but processual: an emergent, dynamically maintained informational structure or mental model arising from the interaction of bodily, neural and environmental processes [19,20,29].

Second, multiple contributions emphasize that the self has an evolutionary and developmental history [11,19,20,29,49], and that a minimal form of selfhood emerges early, with body ownership and sensorimotor agency as important components [19,20,31]. Several authors also employ hierarchical or layered frameworks to distinguish between this minimal form of selfhood, which is characterized as being 'anoetic' (unaware of itself as a self) [59,60], and higher levels of selfhood—self-concept, autobiographical memory, social identity and narrative—which are characterized as 'noetic' (self-aware in the present) or 'autonoetic' (self-aware across time) [12,19,20,29,30,61].

Third, the foundational self–world distinction is considered to be grounded in the physical body [19,31]. Primary self-specifying signals—such as visceral interoceptive dynamics, motor reafference and somatosensory skin boundaries—provide the initial backdrop that separates 'me' from 'not me' [11,12,53]. The construction of the self from this grounding is then an active process that can be characterized computationally as a form of predictive inference [19,20,29,30].

Fourth, the experience of selfhood is described as extending beyond the immediate moment and the physical boundary of the body through temporal integration (spatio-temporal brain dynamics, predictive coding and memory) [20,29,30,55,58] and social engagement (shared agency, dyadic communication, 'we-mode' cognition and moral group identity) [10,13,45,49]. In this context, narratives provide a framework for prediction and social coordination, allowing individuals to present themselves as well-adapted participants in their socio-cultural networks.

Finally, the self is seen as dynamical, that is as a complex system, that arises within and is coupled with states of the body, and through that with the environment [19,29,30]. From this perspective, self-states can be viewed as attractor states, of which some of the most stable become parts of our personal identities.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

5. Reconciling I with me

As noted earlier, William James famously described the self as a duality, writing the following:

'Whatever I may be thinking of, I am always at the same time more or less aware of myself, of my personal existence. At the same time it is I who am aware; so that the total self of me, being as it were duplex, partly known and partly knower, partly object and partly subject, must have two aspects discriminated in it, of which for shortness we may call one the Me and the other the I.' [3, p. 176]

(a) Rethinking self as object

The contributions to the issue are largely in agreement that the 'Me' aspect of self is constituted by different forms of self-knowledge as characterized by Neisser [4–6] as including ecological, interpersonal, temporally extended, conceptual and private forms. However, these new works also show some of the limitations of Neisser's framing.

First, Neisser's ecological self focused on the self as an embodied actor/perceiver, emphasizing exteroceptive and proprioceptive/kinaesthetic aspects of the organism's relation to its environment. In contrast, many contributions here stress the importance of interoception—that is, perceptual access to the organism's internal physiological milieu [11,12,29,53].

Second, where Neisser focused on dyadic face-to-face interactions in defining the interpersonal self, contributions to this issue stress how the interpersonal aspect of self also reflects and integrates the normative and value-laden perspective of the wider cultural group [29,49,55].

Finally, where Neisser's scheme was in part developmentally inspired—the ecological and interpersonal selves precede the temporally extended, conceptual and private forms—this may also require some rethinking. For instance, and as discussed below, Neisser's ecological self explicitly incorporates agency [4], but this may obscure an important developmental progression in the form of agency itself [11,52,62]. In particular, sensorimotor agency can precede the capacity to represent and reason about one's own future actions as an agent. Indeed, developmental and synthetic contributions to this theme issue help to explain how the adult organization of the self, including conceptual forms of body representation, agency and understanding of others, emerges from pre-reflective sensorimotor foundations [11,19,20,29].

(b) Grounding subjective experience

Multiple contributions can also be characterized as grounding James' experiential 'I' in the physically embodied nature of the human condition. Specifically, they emphasize that the continuous stream of non-conceptual interoceptive and somatosensory processing provides the baseline on which higher-level experiences of self are later constructed [11,12,19,29,31,45,58].

The 18th century philosopher Thomas Reid argued that sensation and the feeling of sensation are indistinguishable, writing: 'there is no difference between the sensation and the feeling of it—they are one and the same thing […] in sensation there is no object distinct from the act of the mind by which it is felt.' [63, p. 194]. This idea motivated Stout's account of anoetic consciousness [60], which argued that, in the anoetic case, experience lacks an object distinct from the act of experiencing—sentience is not an awareness of something but is itself the experience. A complementary, modern dynamical perspective treats subjective experience as processual and as emerging from the ongoing interaction of brain, body and environment [64–70]. For example, O'Regan [65] proposes that the feel of a soft object, such as a sponge, consists in mastery of the sensorimotor contingencies (SMCs) associated with 'squishing' the object between your fingers. Extended to interoception, the SMC position suggests that the feel of interoceptive signalling lies not simply in the signals themselves, but in their embedding within the regularities of homeostatic and allostatic processing [18,64]. Similarly, the feel of agency has to do with how sensorimotor schemes are organized and enacted; that is, experiential agency arises first from establishing and maintaining reliable sensorimotor interactions with the environment [66,69].

Kelso [69] discusses the emergence of this early form of anoetic agency from the perspective of dynamical systems theory, characterizing it as a bifurcation in a coupled system caused by positive feedback. Using the example of an infant whose actions move a mobile (attached by a piece of string), he writes, 'The baby's initially spontaneous movements cause the world to change, their perceived consequences have a sudden and sustained amplifying effect on the baby's further actions. This autocatalytic mechanism [a synergetic phase transition] is continuous with our understanding of how biological form develops […]. The root soil of agency […] rests on primal animation, on being alive and moving' [69, p. 492].

More broadly, the emergence of increasingly sophisticated forms of selfhood can be usefully described as involving a series of developmental bifurcations, characterized by the emergence and stabilization of new attractor states [14,19,29,30]. These transitions reorganize the landscape of possible states such that subsequent forms of self-organization constrain or transform the accessibility of earlier forms [19,29,70]. Among other consequences, this helps explain why it is so hard for us as adults to reimagine the experience of our childhood selves.

An important transition is that from anoetic to noetic experience—from being an actor/perceiver to being aware of 'oneself' as that actor/perceiver [11]. For example, in relation to agency, we have noted a progression during infancy from an anoetic form that links sensorimotor capacity with outcomes, to an intentional (noetic) form in which the infant is able to formulate future goals and organize perception and action systems in pursuit of those goals (see also [11,19,29,62,71]). The latter represents a qualitatively different form of agency—rather than merely discovering that 'I can make this happen', the child can represent 'I will do this in order to make that happen'. In dynamical terms, this involves the emergence of a new attractor corresponding

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

not simply to an organism–environment coordination but to an internally maintained prospective state (a goal) that can organize subsequent sensorimotor dynamics. Verschure [29,45] gives these attractor dynamics a control reading: goals are the attractors whose attainment restores the agent's viability, and volition is the covert generation, selection and commitment of trajectories towards them before any action is released. Whilst we have focused here on agency, a similar anoetic to noetic transition likely takes place with respect to body schema [19,72].

As we have noted, from the layered architecture perspective, conceptual selfhood can be understood as emerging from a stack of processes that learn by association and prediction and that act to reduce error [11,19,20,29,42,45]. At a lower level, associative learning may link motivational and interoceptive states with actions and their sensory consequences [19]. At a higher level, predictive models can anticipate how these internal states will change as a consequence of different actions and their outcomes [19]. This constitutes a transition from reactive homoeostatic regulation to predictive, allostatic regulation [29].

This is still all at the anoetic level but the commonalities between these sets of priors and their regularities through time may generate a yet higher set of invariances [30] or what Butz *et al.* call a conceptual frame [20]. This is the noetic or conceptual self that is able to conceive of its own agency and to formulate and pursue specific goals. Through social interaction, including through the acquisition of language, this internal mental construct becomes linked with the shared conceptual spaces of family and wider society [11,13,55]. In this way, notions of self acquired through culture are assimilated to the endogenously generated frame that also adapts to accommodate these shared cultural meanings [29,49,55]. A shorthand for this process is that the developing child, supported by her carers, constructs her own selfhood through a staircase of generative expectations that ultimately come to support an explicit conceptualization of her ownership over sensations, perceptions and actions. With the development of the narrative self, this extends to conceptualizing, and having autonoetic experience of, past and future. Importantly, these anoetic, noetic and autonoetic levels of selfhood, in addition to scaffolding later development, also exist alongside and modulate each other in the adult self.

We have argued above that the experiential 'I', particularly the 'raw feel' of being a self, is grounded in more primitive, non-conceptual processes, including those involving the body's interoceptive and sensorimotor systems. The phenomenal self can therefore be understood as an emergent property of the integrated dynamics of this embodied system. This constitutes a weakly functionalist claim: if subjective experience depends primarily on the organization of these processes rather than on their particular biological substrate, then, in principle, a non-biological entity, such as a physical robot, could possess subjective experience, provided that it instantiated the relevant organization. However, this remains a lively area of debate, with some enactivist [73,74] and biological naturalist positions [75] maintaining that such organization alone may not be sufficient for felt experience and proposing additional constraints, for instance relating to the autonomous and self-maintaining (autopoetic) character of living organisms.

Interestingly, evidence from psychedelic studies seems to support the notion that the core experience of being a self is processual and at least partially abstracted from its bodily and sensorimotor underpinnings. Analysing trip reports from participants using a powerful psychedelic, Day and Zahavi [38] highlight that even when individuals experience a total loss of bodily ownership, bodily agency and spatial location, and report feeling completely 'disembodied', a radically minimal form of self-consciousness—a perspectival 'for-me-ness'—remains intact.

6. Contributions to the theme issue

In the remainder of this introduction, we briefly review each of the contributed articles to the theme issue organized across five thematic areas: (i) development and architecture; (ii) embodiment, memory and self-continuity; (iii) the social self; (iv) therapeutic and health implications; and (v) philosophical and societal issues.

(a) Development and architecture

The first four contributions explore the architectural and developmental foundations of selfhood by integrating insights from developmental psychology, neuroscience and robotics.

Southgate [11] investigates how self-representation emerges during infant development. She argues that the self is not a given but is constructed through developmental processes grounded in sensorimotor embodiment. Southgate highlights how these foundational skills allow an agent to transition from simple sensorimotor loops to more complex, self-referential internal models, providing a framework for understanding the origins of selfhood.

Butz *et al.* [20] explore the core computational components and inductive biases required to create a minimal sense of self for a robot. They propose that the formation of a hierarchical, generative model of the self is driven by active inference processes that allow the agent to interactively infer both its contextual frame and appropriate behaviour within that frame. They discuss experimental implementations in humanoid robots showing how foundational components of self, including body ownership and agency, can develop computationally. This contribution also proposes 'embodied Turing tests' to evaluate the depth of development in artificial self-models.

Manneschi *et al.* [19] describe a layered machine learning framework that models how a self–world distinction emerges during development. Using a simulated humanoid robot, they show how an agent can distinguish its body from its surroundings by learning the dependencies between proprioception and vision during infant-like 'motor babbling'. This foundational capacity then 'scaffolds' a higher layer that learns the continuous dynamics of these signals using neural differential equation

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

models. The study demonstrates that predictive learning occurs more rapidly when the lower-layer self-model is already in place, offering a mechanistic explanation for how the self–world distinction emerges in the human mind/brain.

Verschure [29] argues that volition and agency are emergent properties that arise through 'active decoupling'–when an agent goes beyond being a transducer of stimuli and becomes a source of action governed by counterfactual trajectories, it generates over its own causal model of the world. Verschure describes the Distributed Adaptive Control (DACva) architecture as a mechanistic implementation of this theory that is implemented across five hierarchical layers–somatic, reactive, adaptive, contextual and reflective. By stating the computational principles of DACva as a formalism and grounding each element in identifiable neural circuits, Verschure seeks to construct a falsifiable account of responsible agency in biological and artificial systems, including the self-model that anchors them. Based on these principles, he then proposes the Embodied Situated Responsible Autonomy (ESRA) test as a replacement for Turing's imitation game.

Coming from developmental psychology and developmental robotics, respectively, Southgate and Butz *et al.* both emphasize that self-representation is a constructed property arising from sensorimotor embodiment and predictive inference. Manneschi *et al.* [19] provide a concrete example of this construction through a scaffolded machine learning model where a minimal self–world distinction accelerates later learning of a predictive model. These authors also posit this as an example of a more general strategy for understanding how noetic and autonoetic aspects of the self are constructed upon, and scaffolded by, an anoetic core. Finally, Verschure integrates insights into layered representation and control into a comprehensive control architecture that enables high-level agency and responsible autonomy. Collectively, these contributions support the view that the self is an emergent property of integrated, hierarchical or layered control systems.

(b) Embodiment, memory and self-continuity

The next three contributions shift the focus towards understanding the experience of the continuity of the self through time.

Northoff *et al.* [30] propose a temporal model to explain how extrinsic stimuli are associated with the self. They identify three key mechanisms: temporal matching, integration and differentiation. Temporal matching suggests that external stimuli are linked to the self when their stochastic structure resonates with the brain's own activity. The model frames the self as a spatio-temporal dynamical process, specified across a hierarchy of timescales, and extends these ideas to the debate about artificial selfhood, asking whether AI systems could support the kind of temporally-nested dynamics proposed here for biological selfhood.

Addis *et al.* [58] systematically review the non-mnemonic processes underlying the sense that the self persists through time, described here as diachronic unity or 'phenomenological continuity'. While memory for events or life experiences is often seen as the primary driver of the experience of the continuity of the self, these authors identify five other candidate factors or processes: interoception, embodiment, temporality, life-trajectory stability and nostalgia. They argue that a comprehensive understanding of diachronic unity should take account of these diverse factors. For instance, they highlight how the constancy of bodily signals and the feeling of inhabiting the same body provide a source of continuity even when memory fails.

Ren *et al.* [31] review the role of bodily self-consciousness (BSC) in shaping episodic memory and autonoetic consciousness focusing on four core components—self-identification (body ownership), self-location, first-person perspective and sense of agency—and their experimental manipulation using virtual reality, multisensory conflicts and visuomotor feedback. Classic memory theories state that retrieval is enhanced when the internal or external context matches the encoding environment [76]. Ren *et al.* highlight neuroimaging evidence showing that when BSC components are preserved during encoding, there is significantly higher encoding–retrieval similarity—the reactivation of the same neural patterns during retrieval that were active during encoding. This account therefore extends traditional memory models by showing that BSC provides an experiential bridge between the encoded self and the retrieved self, transforming memories of past events into personally relevant memories of 'my past'.

Where Northoff *et al.* provide a mechanistic account of self-association grounded in the brain's temporal dynamics, Addis complements this by broadening the scope to include non-mnemonic processes like interoception and life-trajectory stability that sustain the sense of self-continuity. Ren *et al.* [31] further integrate these themes by examining the direct link between BSC and the autonoetic capacity to remember one's own past. Together, these contributions frame the self as emerging from integrated spatio-temporal and somatic signals and not exclusively reliant on event memory systems. This view coheres with development evidence that infants experience the self as persisting in time before they can form reliable autobiographical memories [42,77,78].

(c) The social self

The next three contributions illustrate critical dependencies and interactions between sense of self and social context, from the evolutionary roots of social identity to futuristic contexts of joint action during robot teleoperation.

Li *et al.* [49] examine the evolutionary and developmental origins of human social identity, arguing that this is grounded in *moral* identity—adherence to norms and values of a cultural group. Their point of departure is the phenomenon of shared agency in human groups, which they propose to have evolved from individual agency in primate ancestors, to joint agency and then to collective agency. Group foraging played a crucial role in the first step, where interpersonal feedback enabled members of small social groups to formulate and share goals—turning 'I-level' agencies into shared 'we-level' ones. An increase in group size and intergroup competition then facilitated a second extension of (collective) agency to encompass a larger social group with shared culture and norms. According to this view, social identity reflects the emergence of sensitivity to identity markers (clothing, language, culture and so on) alongside a self-reflective capacity to construe oneself as part of a wider collective 'we'.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

While Li and colleagues' paper focuses on the evolutionary origins of social identity, **Vogeley *et al.*** [13] shift the readers' attention to present-day communication, and particularly, its temporal and non-verbal aspects. Following Mead [79], they stress the fundamental relatedness of individuals to others (the 'interdependent self') and the role of others' perceptions in structuring how we see ourselves. In this context, dyadic interactions are analysed via the notion of 'chronemics'—the temporal organization of communication in interaction, evolving in real time with implicit non-verbal signals (e.g. gaze direction and gestures) and continuously driving social exchange. These authors propose that some neurodiverse conditions, including autism, can be explained through differences in sensitivity to chronemics, with a lack of temporal alignment between partners leading to disturbed communication.

Wozniak *et al.* [10] move from the present to the future by addressing the social self in the context of the use of future tools such as humanoid robots. These authors investigate how the brain processes 'we-information', cues that specify the relation between two partners' actions but not the actions themselves, while two people share control of one humanoid robot. Using dual electroencephalogram recording, they show that we-information engages a cascade of perceptual, action-selection, working-memory and response-preparation processes, shifts covert attention towards the partner and speeds joint responses, and that shared control increases self–other overlap with both the partner and the robot. These findings contribute to the understanding of the social self by highlighting the mechanisms of shared control and self–other integration during real-time cooperation. They also demonstrate that interpersonal aspects of self can extend to non-human agents such as humanoid robots.

Collectively, these papers demonstrate that selfhood is, in part, a social mechanism and a cultural construct, negotiated through collaborative agency, shared goals and norms and continuous interaction. All three contributions centre on the transition from isolated individual processing ('I') to shared collective frameworks ('We'), highlighting the relational and extended nature of the human self. Crucially, social others and cultural groups do not merely influence the self; rather, their perspectives are internalized to form part of the core evaluative architecture of the self.

(d) Therapeutic and health implications

The next three contributions highlight the profound therapeutic and health implications of research on the self.

The first two articles identify interoception—the sensing of internal processes such as heartbeat, respiration and digestion—as a foundational process that shapes self-attribution and body image from early development through to adulthood.

Savage and Garfinkel [12] develop a framework for analysing higher-order interoceptive awareness highlighting two processes: interoceptive attribution and appraisal. Interoceptive attribution is concerned with determining the cause of internal sensations, while appraisal ascribes the meaning of those sensations to the self. Drawing on clinical insights and interoception research, these authors suggest that disruptions in these processes are central to various mental health conditions, including anxiety. They demonstrate that this framework offers a structured way to understand how individuals perceive and interpret their bodies and discuss potential targets for clinical interventions aimed at correcting distorted self-perceptions and maladaptive cognitive biases.

von Mohr *et al.* [53] investigate how interoception shapes body image during early adolescence. This critical developmental period is marked by dynamic physiological and social changes that are reflected in how adolescents perceive themselves. These authors present new experimental findings concerning the relationship between interoceptive accuracy and body image dissatisfaction, finding that lower interoceptive accuracy predicts greater dissatisfaction, an effect that becomes more pronounced in the later phases of puberty. By linking internal signals to body image, this research provides insights into the foundations of self-perception and how its disturbances can impact psychological health. The study also suggests that improving interoceptive awareness could help maintain a positive self-image.

Belinger *et al.* [39] explore the potential health benefits of rebalancing self-related processes through an intervention involving psychedelic drugs. Specifically, these authors conducted a randomized, placebo-controlled trial to investigate how a single dose of psilocybin affects self-concept and self-referential processing in patients diagnosed with MDD. The study found that psilocybin led to significant, sustained reductions in depressive symptoms and an increase in positive self-concept two weeks post-administration. Neuroimaging revealed psilocybin-associated modulation of the caudate nucleus, a brain region linked to reward processing and top–down control. These findings support the view that psychedelic-assisted therapy influences multiple levels of self-processing, including subjective self-schemas and their underlying neural correlates.

Together, these contributions illustrate that a deeper understanding of the self's physiological and cognitive components can lead to more effective, targeted treatments for mental health conditions and to the improvement of psychological well-being. Together with other contributions [13,30], they highlight the growing recognition that many psychiatric conditions are best understood as disorders of the sense of self.

(e) Philosophical and societal implications

The final three contributions examine some core philosophical and societal concerns related to selfhood, ranging from identity construction through to the nature of free will.

Schechtman [55] defends theories of the 'narrative self' against the criticism that self-narratives are aesthetic artefacts that distort life's 'messy' reality. Focusing on how individuals harness episodic memory to produce self-narratives that organize

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

experience, Schechtman argues that self-narration is not a choice but a developmental stage. Specifically, the construction of narratives helps individuals to constitute their 'mature human selves' by tracking their lives within cultural constraints. From this perspective, even when narratives are partially inaccurate, they can be beneficial by providing a stable framework for prediction and social coordination.

Day and Zahavi [38] present a critique of the claim that psychedelic-induced ego-dissolution results in 'totally selfless' conscious states. Based on a thematic analysis of trip reports, they argue that while narrative and embodied senses of self can be lost, a 'radically minimal' form of self-consciousness—characterized as 'for-me-ness'—is still present. The authors maintain that extreme ego-dissolution is extraordinary not because the subject is absent, but because it exemplifies a pure form of subjectivity maintaining a first-person point of view.

Finally, in a second contribution to this theme issue, **Verschure** [45] proposes replacing metaphysical questions about free will with the control problem of 'responsible autonomy'—the functional state in which volition and agency are constrained by normative self-governance, so that an agent can be held accountable. Verschure specifies six capacities: active decoupling, endogenous drive, normative self-governance, diachronic integrity, leeway and experiential ownership, derived from the need for agents to survive in social worlds defined by hidden intentional states. Each capacity is argued to be necessary by deletion, and the six capacities become architectural requirements that a theory of agency, biological or artificial, must meet, and that the DACva architecture [29] formalizes. Verschure argues that the notion of free will is thereby naturalized as the control interface that embodied agents have evolved to navigate in these social worlds.

Schechtman and Day and Zahavi both probe the fundamental structure of the self, examining, respectively, its high-level narrative functions and its most minimal experiential foundations. Both argue that the self—whether narrative or minimal—is a ubiquitous and perhaps inescapable organizing principle of human existence. Verschure focuses on the question of what a self must be for it to be held responsible, arguing that this can be posed as a control problem with testable requirements and so brought under governance. Together, these works show that a scientific understanding of the self can shed light on the human condition, both in terms of the fundamental nature of selves and our responsibilities towards each other.

7. Open questions and future work

The contributions to this theme issue raise a range of open questions about the nature of the self and promising avenues for new research. Here we highlight several broad topics where we believe progress could be made.

(a) Understanding the emergence of the self

A central mechanistic challenge is to explain how pre-reflective, anoetic sensorimotor foundations—such as interoceptive streams [11], scale-free resting-state dynamics [30] and motor babbling loops [19]—are progressively transformed and integrated into higher-order noetic and autonoetic structures. Future synthetic models should demonstrate how minimal self–world boundaries can scaffold the acquisition of prospective counterfactual reasoning, complex problem-solving and a fully realized Theory of Mind [19,20,29]. Understanding how multi-scale temporal dynamics—from millisecond sensorimotor contingencies to multi-year autobiographical timelines—organize the self-system remains an important frontier for neuroimaging, developmental psychology and cognitive robotics [11,20,30,31,58].

(b) The possibility of synthetic selves

The prospect of constructing artificial selves brings urgent ethical and societal concerns. As agentic AIs and autonomous robotic systems become increasingly integrated into human society, we face a governance mismatch between legal frameworks that presuppose moral agency and scientific ontologies that risk reducing selfhood to an illusion [45].

Because large language models can fluidly 'role-play' subjective first-person narratives without physical embodiment or unmediated world engagement [80], validating whether a synthetic system possesses a genuine self requires rigorous benchmarks. These could include embodied Turing tests [20], alongside formal audits such as the ESRA test [29], which assesses any agent, biological or artificial, on six explicit control capacities, including active decoupling, endogenous drive and normative self-governance. Verschure [29] applies the ESRA test to reports of two 2026 security incidents involving foundation AI models concluding that, in both cases, the systems proved deficient on at least three of the six measures (the remaining three being undecided owing to lack of information). He concludes that these incidents reflect a failure of specification, containment and monitoring on the part of the companies that developed the models rather than the emergence of a new kind of intrinsically motivated and wilful AI.

This finding demonstrates that a key ethical priority is to avoid anthropomorphic inflation—mistaking the generative linguistic fluency of AI models for subjective experience, first-hand agency or moral patiency [29,47]. Ungrounded AI agents that lack allostatic drives or internalized normative constraints cannot be held genuinely accountable, and to misattribute these qualities risks dangerous buck-passing, excuses human negligence and will not solve the most pressing AI safety problems [29,45]. Conversely, as and when synthetic architectures approach true normative self-governance, society must establish clear ethical boundaries regarding machine rights and moral status [81] and the limits of synthetic agency [45].

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

(c) The self and psychological well-being

A central insight emerging from this theme issue is that many psychiatric conditions are fundamentally disorders of the self, arising from disruptions across its layered sensorimotor, cognitive and narrative subsystems [12,13,39,53]. Reframing clinical conditions through this lens offers novel targets for therapeutic intervention. For instance, clarifying how interoceptive signals are consciously processed reveals that higher-order attribution and appraisal dictate whether visceral arousal translates into adaptive emotional regulation or clinical anxiety [12]. Clinical interventions that modulate self-referential processing—such as psilocybin-assisted therapy—demonstrate that rebalancing attentional engagement away from ruminative self-schemas towards an externally directed focus leads to sustained improvements in positive self-concept and reductions in depressive severity [39].

This speaks to a more general view of therapeutic treatments as a means of adjusting the attractor dynamics of the self system away from regimes that negatively impact psychological well-being. Disorders of self can arise when the system's attractor landscape becomes altered—through excessive stability, excessive instability, inappropriate attractors or abnormal transitions between partially competing self-models [82]. To make useful progress requires that we better specify the relevant state variables, identify proposed attractors and make predictions about perturbation and recovery. This could be one of the key benefits of developing synthetic models of the self system [21,23].

(d) The future of the self

As human existence becomes increasingly intertwined with digital and synthetic environments, technologies, such as virtual reality, telepresence robotics and agentic artificial intelligence, are fundamentally altering the spatial, temporal and social boundaries of the self [10]. Immersive virtual reality and robotic telepresence demonstrate the malleability of BSC, allowing body ownership and first-person perspective to decouple from the physical organism [31,35]. When individuals jointly teleoperate robotic artefacts, shared control increases self–other overlap with human and robotic partners and engages we-mode processing at the neural level [10].

With the rapid deployment of agentic AI systems and autonomous systems such as self-driving vehicles, humans are increasingly delegating aspects of their own agency to artificial others. Persistent delegation of action selection to autonomous artificial agents could alter the dynamical processes through which humans maintain a sense of agency and authorship, particularly if delegation reduces opportunities for self-generated action and action–outcome feedback [83]. A better understanding of human–machine shared agency and of the relationship between sense of agency, motivation and responsibility should help guide the development of more useful and beneficial assistive technologies and agentic systems.

8. Conclusion

By bringing together analytic insights from philosophy, psychology and neuroscience alongside synthetic methods in robotics and AI, this theme issue demonstrates the potential of an integrative science of the self. Rather than viewing the self as a localized substance, or dismissing it as an illusion, the contributed papers in this collection converge towards a multilayered, dynamical systems perspective. In this framework, the self emerges as the consequence of an embodied dynamical and informational system continuously distinguishing, regulating and modelling itself through processes that sustain and explain its own existence. This view reunifies William James' classical distinction between the experiencing subject ('I') and the known object ('Me'). Grounded initially in pre-reflective bodily and interoceptive signals that delineate the foundational self–world boundary, the self extends across time and space through prediction, memory, spatio-temporal brain dynamics and narrative construction. Crucially, human selfhood is profoundly relational, the perspectives of social others and cultural norms being internalized into its core architecture.

We believe that bridging between analytic and synthetic approaches deepens our understanding of human selfhood while offering new frameworks for understanding the diversity of human selves and governing responsible artificial agency. The synthetic approach can serve a dual purpose: it acts as a diagnostic engine for testing biological theories of mind while ensuring that as we build increasingly autonomous artefacts, we remain anchored in a naturalistic, human-centred understanding of responsibility, dignity and selfhood.

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

Data accessibility. This article has no additional data.

Declaration of AI use. Generative AI tools were used in the preparation of this introduction to assist with language editing, drafting and refinement. The authors take full responsibility for the content, interpretation and accuracy of the text.

Authors' contributions. T.J.P.: writing—original draft, writing—review and editing; A.W.: writing—review and editing; S.N.G.: writing—review and editing; P.F.M.J.V.: 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. This theme issue was put together by the Guest Editor team under supervision from the journal's Editorial staff, following the Royal Society's ethical codes and best-practice guidelines. The Guest Editor team invited contributions and handled the review process. Individual Guest Editors were not involved in assessing papers where they had a personal, professional or financial conflict of interest with the authors or the research described. Independent reviewers assessed all papers. Invitation to contribute did not guarantee inclusion.

T.J. Prescott is a director and shareholder of the UK company Consequential Robotics Ltd. P.F.M.J. Verschure is a shareholder of the ES company Eodyne Systems s.l.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

Funding. This theme issue arose from a Hooke Scientific Discussion Meeting hosted by the Royal Society in London on 19–20 May 2025. The meeting received additional support from European Innovation Council (EIC) Pathfinder projects CAVAA (project no. 101071178) and VALAWAI (project no. 101070930). Both projects are part of the EIC's 'Awareness Inside' programme. Support from the CAVAA project was partly provided by Innovate UK under the United Kingdom's funding guarantee scheme for Horizon Europe.

Acknowledgements. We are grateful to the early-career researchers who presented posters and spotlight talks at the meeting and to the discussants on our lived-experience panel for their thoughtful contributions on neurodiversity and sense of self. We also thank all participants who attended in person or online. We are especially grateful to the Scientific Meetings Team at the Royal Society, particularly Harry Cutmore, for organizing a smoothly run and technically flawless event, and to Helen Eaton of the *Philosophical Transactions* team for her invaluable editorial support. The articles in this theme issue have also benefited greatly from the insightful and thorough reviews provided by our international reviewers.

References

1. Descartes R. 1641 *Meditationes de prima philosophia, in qua dei existentia et animae a corpore distinctio demonstrantur*. Paris, France: Michael Soly.
2. Chamberlain C. 2020 What am I? Descartes's various ways of considering the self. *J. Mod. Phil.* **2**, 2. (doi:10.32881/jomp.30)
3. James W. 1892 *Psychology: the briefer course*. New York, NY: Harper. (doi:10.1037/11630-000)
4. Neisser U. 1988 Five kinds of self-knowledge. *Philos. Psychol.* **1**, 35–59. (doi:10.1080/09515088808572924)
5. Neisser U. 1995 Criteria for an ecological self. In *The self in infancy: theory and research*(ed. P Rochat), pp. 17–34. Amsterdam, The Netherlands: Elsevier.
6. Neisser U. 1994 *The perceived self*. Cambridge, UK: Cambridge University Press.
7. Dennett DC. 1992 The self as a center of narrative gravity. In *Self and consciousness: multiple perspectives* (eds FS Kessel, PM Cole, DL Johnson, MD Hakel), pp. 103. Oxford, UK: Psychology Press. (doi:10.5209/rev-ASEM.2013.v46.42862)
8. Schechtman M. 2011 The narrative self. In *The Oxford handbook of the self*(ed. S Gallagher), pp. 394–416. Oxford, UK: Oxford University Press. (doi:10.1093/oxfordhb/9780199548019.003.0018)
9. Schechtman M. 2019 *The constitution of selves*. New York, NY: Cornell University Press. (doi:10.7591/9781501718380)
10. Wozniak M, Haxhi R, De Tommaso D, Wykowska A. 2026 Two minds, one body: neural correlates of processing of we-information during shared control over a humanoid robot. *Phil. Trans. R. Soc. B* . **381**, 20250163. (doi:10.1098/rstb.2025.0163)
11. Southgate V. 2026 How does the developing mind construct a self? *Phil. Trans. R. Soc. B* **381**, 20250158. (doi:10.1098/rstb.2025.0158)
12. Savage H, Felsenheimer A, Smekal V, Garfinkel S. 2026 Making sense of sensation: a model of interoceptive attribution and appraisal with clinical applications. *Phil. Trans. R. Soc. B* **381**, 20250161. (doi:10.1098/rstb.2025.0161)
13. Vogeley K, Wünsche M, Jording M. 2026 The social self: integration of multimodal communication signals in dyadic interactions. *Phil. Trans. R. Soc. B* **381**, 20250164. (doi:10.1098/rstb.2025.0164)
14. Prescott TJ, Camilleri D. 2019 The synthetic psychology of the self. In *Cognitive architectures*(eds MI Aldinhas Ferreira, J Silva Sequeira, R. Ventura), pp. 85–104. Cham, Switzerland: Springer International Publishing. (doi:10.1007/978-3-319-97550-4_7)
15. Damasio AR. 2010 *Self comes to mind: constructing the conscious brain*. London, UK: Vintage Books.
16. Metzinger T. 2009 *The ego tunnel: the science of the mind and the myth of the self*. New York, NY: Basic Books.
17. Gallagher S. 2013 A pattern theory of self. *Front. Hum. Neurosci.* **7**, 1–7. (doi:10.3389/fnhum.2013.00443)
18. 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)
19. Manneschi L, Jimenez-Rodriguez A, Camilleri D, Prescott T. 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. (doi:10.1098/rstb.2025.0153)
20. Butz M, Hafner V, Lanillos P. 2026 Building a robotic self. *Phil. Trans. R. Soc. B* **381**, 20250166. (doi:10.1098/rstb.2025.0166)
21. Prescott TJ, Vogeley K, Wykowska A. 2024 Understanding the sense of self through robotics. *Sci. Robot.* **9**, eadn2733. (doi:10.1126/scirobotics.adn2733)
22. Prescott TJ. 2021 Robot self. In *Encyclopedia of robotics*. (eds MH Ang, O Khatib, B SicilianoBerlin, Germany: Springer. (doi:10.1007/978-3-642-41610-1_205-1)
23. Hafner VV, Loviken P, Pico Villalpando A, Schillaci G. 2020 Prerequisites for an artificial self. *Front. Neurorobot.* **14**, 5. (doi:10.3389/fnbot.2020.00005)
24. 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)
25. Hohwy J, Michael J. 2017 Why should any body have a self? In *The subject's matter: self-consciousness and the body* (eds F de Vignemont, A Alsmith). Cambridge, MA: MIT Press. (doi:10.7551/mitpress/10462.003.0043)
26. Newen A. 2018 The embodied self, the pattern theory of self, and the predictive mind. *Front. Psychol.* **9**, 2270. (doi:10.3389/fpsyg.2018.02270)
27. Morales FX. 2025 Self/other: guiding distinction for the social theory of the self. *Curr. Sociol.* **73**, 595–612. (doi:10.1177/00113921251316683)
28. Hards E, Rathbone CJ, Ellis JA, Reynolds S. 2024 'What is the self anyway?' towards a more parsimonious conceptualisation of the self: a review. *New Ideas Psychol.* **74**, 101080. (doi:10.1016/j.newideapsych.2024.101080)
29. Verschure P. 2026 The architecture of responsible autonomy in biological and artificial systems: the distributed adaptive control of volition and agency. *Phil. Trans. R. Soc. B* **381**, 20250156. (doi:10.1098/rstb.2025.0156)
30. Northoff G, Shen X, Avramescu A, Sui J. 2026 From rest to task—temporal mechanisms of self-association. *Phil. Trans. R. Soc. B* **381**, 20250346. (doi:10.1098/rstb.2025.0346)
31. Ren Q, Gui-Bordeianu A, Blanke O. 2026 From body to memory: the role of bodily self-consciousness in episodic memory and autonoetic consciousness. *Phil. Trans. R. Soc. B* **381**, 20250167. (doi:10.1098/rstb.2025.0167)
32. 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)
33. Levorsen M, Aoki R, Matsumoto K, Sedikides C, Izuma K. 2023 The self-concept is represented in the medial prefrontal cortex in terms of self-importance. *J. Neurosci.* **43**, 3675–3686. (doi:10.1523/JNEUROSCI.2178-22.2023)
34. Zhao K, Dang J, Gu J, Fu X, Haggard P. 2025 Cognitive mechanisms underlying sense of agency: meta-analytic reviews of behavioral and neuroimaging studies. *Psychol. Bull.* **151**, 1307–1335. (doi:10.1037/bul0000497)

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

35. 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)
36. Letheby C, Gerrans P. 2017 Self unbound: ego dissolution in psychedelic experience. *Neurosci. Conscious.* **2017**, nix016. (doi:10.1093/nc/nix016)
37. Deane G. 2020 Dissolving the self. *Phil. Mind Sci.* **1**, 1–27. (doi:10.33735/phimisci.2020.i.39)
38. Day JK, Zahavi D. 2026 Psychedelic ego-dissolution: a philosophical and qualitative critique. *Phil. Trans. R. Soc. B* **381**, 20250154. (doi:10.1098/rstb.2025.0154)
39. Belinger L. *et al.* 2026 Changes in self-related cognitive processes and depressive symptom severity following a single, moderate dose of psilocybin in depressed patients. *Phil. Trans. R. Soc. B.* **381**, 20250168. (doi:10.1098/rstb.2025.0168)
40. 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: OUP. (doi:10.1093/oso/9780199674923.003.0002)
41. Verschure PF. 2016 Synthetic consciousness: the distributed adaptive control perspective. *Phil. Trans. R. Soc. B* **371**, 20150448. (doi:10.1098/rstb.2015.0448)
42. 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)
43. Lanillos P, Pages J, Cheng G. 2020 Robot self/other distinction: active inference meets neural networks learning in a mirror. *Front. Artif. Intell. Appl.* **325**, 2410–2416. (doi:10.3233/FAIA200372)
44. 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)
45. Verschure PFMJ. 2026 The six capacities of responsible autonomy: naturalising free will in biological and artificial systems. *Phil. Trans. R. Soc. B.* **381**, 20260001. (doi:10.1098/rstb.2026.0001)
46. Jékely G, Godfrey-Smith P, Keijzer F. 2021 Reafference and the origin of the self in early nervous system evolution. *Phil. Trans. R. Soc. B* **376**, 20190764. (doi:10.1098/rstb.2019.0764)
47. Prescott TJ. *The synthetic self*. https://aeon.co/ (accessed 12 January 2026).
48. Dennett DC. 1989 The origin of selves. *Cogito* **3**, 163–173. (doi:10.5840/cogito19893348)
49. Li L, Grüneisen S, Tomasello M. 2026 Evolutionary foundations of social identity. *Phil. Trans. R. Soc. B.* **381**, 20250157. (doi:10.1068/rstb.2025.0157)
50. 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)
51. Amsterdam B. 1972 Mirror self-image reactions before age two. *Dev. Psychobiol.* **5**, 297–305. (doi:10.1002/dev.420050403)
52. Southgate V. 2024 The origins and emergence of self-representation. *Annu. Rev. Dev. Psychol.* **6**, 109–131. (doi:10.1146/annurev-devpsych-120621-025747)
53. von Mohr M, Donaghy R, Ainley V, Moore I, Agyemang A, Tsakiris M. 2026 Tuning in: how interoception shapes body image across early adolescence. *Phil. Trans. R Soc. B* **381**, 20250159. (doi:10.1098/rstb.2025.0159)
54. Fivush R, Haden CA. 2003 *Autobiographical memory and the construction of the narrative self: developmental and cultural perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
55. Schechtman M. 2026 Defending the narrative self. *Phil. Trans. R. Soc. B.* **381**, 20250155.
56. Caggiano P, Jehkonen M. 2018 The 'neglected' personal neglect. *Neuropsychol. Rev.* **28**, 417–435. (doi:10.1007/s11065-018-9394-4)
57. Clark IA, Maguire EA. 2016 Remembering preservation in hippocampal amnesia. *Annu. Rev. Psychol.* **67**, 51–82. (doi:10.1146/annurev-psych-122414-033739)
58. Addis DR, Tippett L, Davis O. 2026 Towards a comprehensive understanding of the processes underlying phenomenological continuity. *Phil. Trans. R. Soc. B.* **381**, 20250162. (doi:10.1098/rstb.2025.0162)
59. Panksepp J. 1998 The periconscious substrates of consciousness: affective states and the evolutionary origins of the self. *J. Conscious. Stud.* **5**, 566–582.
60. Stout G. 1896 *Analytical Psychology,* **vol. 1.** Aberdeen, UK: Aberdeen University Press.
61. 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)
62. Pacherie E. 2026 *Agency*. Elements in Philosophy of Mind. Cambridge, UK: Cambridge University Press. (doi:10.1017/9781009376310)
63. Reid T. 1785/1863 *Essays on the Intellectual Powers of Man*. Cambridge, MA: MIT.
64. Buhrmann T, Di Paolo EA, Barandiaran X. 2013 A dynamical systems account of sensorimotor contingencies. *Front. Psychol.* **4**, 285. (doi:10.3389/fpsyg.2013.00285)
65. O'Regan K. 2011 *Why red doesn't sound like a bell: understanding the feel of consciousness*. Oxford, UK: Oxford University Press.
66. Buhrmann T, Di Paolo E. 2017 The sense of agency – a phenomenological consequence of enacting sensorimotor schemes. *Phenomenol. Cogn. Sci.* **16**, 207–236. (doi:10.1007/s11097-015-9446-7)
67. O'Regan JK, Noe A. 2001 A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* **24**, 939–973. (doi:10.1017/s0140525x01000115)
68. O'Regan JK. 2023 How voluntary control over information and body movements determines "what it's like" to have perceptual, bodily, emotional and mental experiences. *Front. Psychol.* **13**, 1061930. (doi:10.3389/fpsyg.2022.1108279)
69. Kelso JAS. 2016 On the self-organizing origins of agency. *Trends Cogn. Sci.* **20**, 490–499. (doi:10.1016/j.tics.2016.04.004)
70. Thelen E, Smith LB. 1994 *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.
71. Delafield-Butt JT, Gangopadhyay N. 2013 Sensorimotor intentionality: the origins of intentionality in prospective agent action. *Dev. Rev.* **33**, 399–425. (doi:10.1016/j.dr.2013.09.001)
72. Rochat P, Botto SV. 2021 From implicit to explicit body awareness in the first two years of life. In *Body schema and body image: new directions* (eds Y Ataria, S Tanaka, S Gallagher), pp. 181–193. Oxford, UK: Oxford University Press.
73. Di Paolo E, Buhrmann T, Barandiaran X. 2017 *Sensorimotor life: an enactive proposal*. Oxford, UK: Oxford University Press. (doi:10.1093/acprof:oso/9780198786849.001.0001)
74. Thompson E. 2005 Sensorimotor subjectivity and the enactive approach to experience. *Phenomenol. Cogn. Sci.* **4**, 407–427. (doi:10.1007/s11097-005-9003-x)
75. Seth AK. 2025 Conscious artificial intelligence and biological naturalism. *Behav. Brain Sci.* **49**, 1–42. (doi:10.1017/S0140525X25000032)
76. Smith SM, Vela E. 2001 Environmental context-dependent memory: a review and meta-analysis. *Psychon. Bull. Rev.* **8**, 203–220. (doi:10.3758/BF03196157)
77. 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.
78. Nelson K. 1993 The psychological and social origins of autobiographical memory. *Psychol. Sci.* **4**, 7–14. (doi:10.1111/j.1467-9280.1993.tb00548.x)
79. Mead GH. 1934 *Mind, self, and society: the definitive edition*. Chicago, IL: University of Chicago Press.

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026

80. Shanahan M, McDonell K, Reynolds L. 2023 Role play with large language models. *Nature* **623**, 493–498. (doi:10.1038/s41586-023-06647-8)
81. Gunkel D. 2018 *Robot rights*. Cambridge, MA: MIT Press.
82. Scheffer M *et al.* 2024 A dynamical Systems view of psychiatric disorders—theory: a review. *JAMA Psychiatry* **81**, 618–623. (doi:10.1001/jamapsychiatry.2024.0215)
83. Elettari J, Prescott TJ, Vogeley K, Wykowska A. 2026 The quest for the optimal level of sense of agency in the case of human-operated, AI-enabled robotics. *Curr. Opin. Behav. Sci.* **71**, 101674. (doi:10.1016/j.cobeha.2026.101674)

Downloaded from royalsocietypublishing.org/rstb/article-pdf/doi/10.1098/rstb.2025.0152/6177098/rstb.2025.0152.pdf
by guest
on 13 November 2026