
Book Section:
Prescott, T.J. orcid.org/0000-0003-4927-5390 and Camilleri, D. (2019) The synthetic psychology of the self. In: Ferreira, M.I.A., Sequeira, J.S. and Ventura, R., (eds.) Cognitive Architectures. Intelligent Systems, Control and Automation: Science and Engineering (94). Springer Nature , pp. 85-104. ISBN 9783319975498

https://doi.org/10.1007/978-3-319-97550-4_7

To appear in: Cognitive Architectures, João Sequeira, Rodrigo Ventural and Isabel Ferraira (eds). Berlin: Springer Verlag.

The Synthetic Psychology of the Self

Tony J Prescott and Daniel Camilleri

The University of Sheffield and Sheffield Robotics

Abstract Synthetic psychology describes the approach of "understanding through building" applied to the human condition. In this chapter, we consider the specific challenge of synthesizing a robot "sense of self". Our starting hypothesis is that the human self is brought into being by the activity of a set of transient self-processes instantiated by the brain and body. We propose that we can synthesize a robot self by developing equivalent sub-systems within an integrated biomimetic cognitive architecture for a humanoid robot. We begin the chapter by motivating this work in the context of the criteria for recognizing other minds, and the challenge of benchmarking artificial intelligence against human, and conclude by describing efforts to create a sense of self for the iCub humanoid robot that has ecological, temporally-extended, interpersonal and narrative components set within a multi-layered model of mind.

Alan Turing, one of the founders of computer science, once suggested that there were two paths to human-level Artificial Intelligence (AI)—one through emulating the more abstract abilities of the human mind, such as chess playing, the other, much closer to the spirit of this book, by providing a robot with "the best sense organs that money can buy, and then teach[ing] it to understand and speak English. This process could follow the normal teaching of a child" [66, p.460]. Turing was noncommittal about which approach would work best and suggested we try both. Two-thirds of a century after Turing, as different AIs battle between themselves to be the world's best at chess [59], it is clear that the first approach has been spectacularly successful at producing some forms of machine intelligence, though not at emulating or approaching "general intelligence"—the

wider intellectual and cognitive capacities of our species.[1] Enthusiasm for Turing's second approach has therefore re-emerged and is continuing to grow.

1 Beyond the Turing test

Even more famously, and in the same paper [66], Turing also suggested a way of deciding whether a machine could think in the form of an "imitation game." In what is now universally known as the "Turing test", a judge is asked to distinguish between a human and a machine based on written communication alone. In devising the test, Turing explicitly sought to avoid defining thinking in terms of unobservables, for example, operations of the mind. Instead, he argued that we should focus on behavioral phenomena, such as the ability to conduct a conversation that, in a human, would be recognized as requiring thinking. The design of the Turing test is intended to create an unbiased way of comparing a machine with a man or woman, since there are no extraneous clues, such as appearance or tone of voice, to reveal which is which. Since 1991, an annual competition, the Loebner prize, has sought to evaluate the ability of AIs to pass tests based on Turing's proposal—a prize of $100,000 stands on offer to the first AI to be consistently mistaken for an adult human following an extended and open-ended conversation.
In *Ex Machina,* the 2015 science fiction movie about future AI, Nathan Bateman, the fictional inventor of Ava, a new kind of humanoid robot, proposes an alternative to the Turing Test, in which "the real test is to show you that she [Ava] is a robot; then see if you still feel she has consciousness."[2] What we might call the "Garland test", af-

[1] By this, we mean the cluster of different but overlapping intellectual/cognitive faculties that make humans adaptive, flexible sociotechnical animals. Howard Gardner's [21] "multiple intelligences" view provides a good guide to this broader notion of human cognition. Attempts to create machine intelligence of this more multi-faceted form are increasingly discussed under the label *Artificial General Intelligence* (AGI) (e.g., [22]), hence we are using the phrase "general intelligence" rather than Gardner's multiple intelligences.

[2] Nathan Bateman to Caleb Smith about the humanoid robot "Ava" he has created, from the original movie script for *Ex Machina* (2015) by Alex Garland.

ter the writer of *Ex Machina*, Alex Garland,[3] is arguably a tougher challenge than the original test devised by Turing—there is no question of whether you are speaking to a robot or a human; the witness you are interrogating is clearly a machine. Yet, like Caleb Smith, the young programmer whom Nathan chooses to interact with his robot, you might feel compelled by the robot's ability to converse and behave in a life-like way to view this machine as having a mind of its own.

It is worth noting that Turing intended his test as a way of deciding whether a machine could think, and not whether a machine has consciousness. Indeed, Turing writes, "I do not think these mysteries [about consciousness] necessarily need to be solved before we can answer the question with which we are concerned in this paper [whether a machine can think]" ([66], p. 447). However, many commentators have considered the test to be about consciousness, for example, John Searle, in describing the Chinese Room, a thought experiment predicated on the Turing test, rephrases Turing's question "can a machine think?" as "can a machine have conscious thoughts?" ([57], p. 20). The Chinese Room is intended to demonstrate that a machine could pass the Turing test in Chinese *without* understanding Chinese. Turing might possibly have agreed. For Searle, and others, thoughts have to come from conscious minds in order to be actual thoughts (to be "about" something), whereas for Turing, it was enough for a system to generate the right kind of behavior to be considered as thinking; consciousness was something else.

Other forms of Turing test have also been proposed by Steven Harnad [23, 24], who has suggested a hierarchy of Turing tests: Level T1 is a narrow AI, for instance, one that can prove mathematical theorems or is exceptional at chess. T2, the original test, demonstrates what Harnad calls "pen-pal" level indistinguishability by emulating human linguistic capacity. T3, the "total Turing test", requires that the robot is capable of emulating human language *and action*, but need not be made of biological stuff or otherwise constrained to match a particular internal structure. For Harnad, T3 is

---

[3] The suggestion that we call this the Garland test has also been made by Murray Shanahan, one of the scientific advisors on *Ex Machina*.

the level at which we judge other people, the point at which symbolic computation becomes "grounded" in the external world, and therefore the correct level at which to judge whether a machine has conscious thoughts.[4] Harnad also describes, but rejects as too stringent, a level T4—detailed biological indistinguishability—as might be required by some anti-functionalist stances.

One of the more intriguing ideas in *Ex Machina* is that we are left unsure, at the end, as to whether the robot, Ava, has a mind similar to ours or whether it is, instead, an alien and devious AI that is able to emulate and deceive humans when this serves its purposes. Does this ending suggest a challenge to Harnad's proposal for a T3 Turing test or, indeed, for the Garland test (which is a variant of that test)? Harnad [24] admits that the T3 test is under-constrained in emulating *how* people think, but like Turing, he is comfortable with that; for Harnad, succeeding in the T3 test is evidence enough of grounded (and conscious) thoughts. However, what if we want to get closer to understanding the mind, or to build a machine that actually does think like a human? The evidence from Chess and Go is that machines can exceed human experts at these intellectual challenges without matching the way in which people play either game. Similarly, T3 equivalence could give us grounded symbols, but without further resolving how human minds work.

But perhaps we can get closer to human general intelligence without going all the way to T4 equivalence. Specifically, suppose we add the constraint of having a human-like *cognitive architecture* in addition to matching human symbolic and robotic capacity. If we can match both the behavior and the architecture of mind, then there is a greater likelihood that our AI will not only act like us but also think like us. Following the scheme of Harnad's test hierarchy, we might call this level T3.5.

---

[4] It has been suggested that Harnad's T2 level cannot be achieved without first building T3 to achieve symbol-grounding [25]. Going directly to T2 is nevertheless a theoretical possibility, even if it might prove impossible to achieve without a contribution from robotics.

2 Robotics as synthetic psychology

Based on this line of reasoning, I have, for the past seven years, been involved in various projects concerned with the development of aspects of general intelligence for humanoid robots. This work builds on the above premise that we can seek to create an artificial mind that is similar to our own by emulating human linguistic and robotic capacity and by employing a cognitive architecture that has been reverse-engineered from findings in psychology and neuroscience. The hope is that we can make significant progress without having to concern ourselves with all of the T4-level detail. The long-term goal is to build a machine that can pass the Garland test whilst being sufficiently biomimetic in design that we can credibly argue that its "mental states" are analogous to human mental states in an interesting way.

This goal can also be seen as belonging to the sub-discipline of *synthetic psychology*, an enterprise within the cognitive sciences named after Valentino Braitenberg's inspirational book *Vehicles: Experiments in Synthetic Psychology* [74], which advocates that we build artificial creatures as a path to understanding the brains and behavior of biological organisms. This "understanding through building" approach also forms a core principle of the emerging field of *Living Machines* [47].[5] Within robotics, there is a growing group of researchers interested in this challenge, indeed, when we add in developmental constraints, this approach to reverse-engineering the human converges within the emerging field of developmental robotics (e.g., [13, 34]).

So, what should the ambition of a synthetic psychologist be in building a human-like machine? For many philosophers and cognitive scientists, even some roboticists, the Holy Grail is to understand and recreate human consciousness. While this ambition is attractive, it

---

5 This idea also follows in the footsteps of many others. For example, the eighteenth century Neapolitan philosopher Giambattista Vico, who wrote "*verum et factum reciprocantur seu convertuntur* [the true is precisely what is made]", and the 20th century physicist Richard Feynman, whose office blackboard on the day he died held the message, "what I cannot create I do not understand".

suffers from two serious drawbacks. First, the difficulty of deciding what consciousness is, and second, the challenge of measuring subjective first-person phenomena using a third-person approach (the tools of science).[6] For this reason, we have chosen not to make consciousness a target of our synthetic psychology research, preferring instead a (hopefully) more tangible phenomenon—to construct a robot with a "sense of self" [48]. Perhaps we will find that we cannot completely disentangle self from consciousness, but even so, by understanding the broader nature of self, we may be able to see more clearly what, if anything, is still left to explain about first-person experience.

3 Defining and deconstructing the self

Some might baulk at the thought of trying to synthesize the self without directly addressing consciousness, others, following David Hume [28], may consider that there is little to be assembled in a synthetic self beyond a stream of perceptions. But there is an interesting third way. For instance, writers such as the psychologists Susan Blakemore [9] and Bruce Hood [27], the cognitive scientist Douglas Hofstadter [75], the architect Chris Abel [1] and the philosopher Thomas Metzinger [37] have argued that the self as we conventionally imagine it is an illusion, but that, nevertheless, there is something there to be understood. For Blakemore, it is a complex of memes, for Hood, an internal simulation, for Hofstadter, a "strange loop", for Abel, a "field of being" that can extend outside the body[7],

---

6 There are multiple measures of so-called "correlates of consciousness", Giulio Tononi's Φ [63], a measure of information integration, being one of the better-known ones. The problem is that there is no way to be sure that an organism or machine that scores highly on any such measure is actually experiencing consciousness. This is known as the "other minds" problem in philosophy. For Turing [66], this was part of the reason to devise a behavioral test for the existence of machine thought and to leave the challenge of consciousness to others.

7 Abel's "field of being" view stems from Maurice Merleau-Ponty's [36] phenomenology and his insistence on the centrality of the experience of the body. Studies in cognitive neuroscience, such as those of the "rubber hand" illusion (see [10]), support Merleau-Ponty's proposal that the sense of the body/self can extend into objects and the world. With virtual reality systems and telepresence robots, it is now possible to experimentally manipulate the sense of a virtual body, or of a physically remote robot body, and the associated feelings of immersion or "presence",

and for Metzinger, a meta-representation (amongst other things). Thus, while for Blakemore, the self is a construct, for Hood, Hofstadter, Abel and Metzinger, the self is also a process, or set of processes, some of which may be representational and reflective, that arise in the brain and body. The proposal we are seeking to investigate is similar: that the sense of self can be emulated by a set of definable and buildable processes that can be situated in some suitably configured robot.

The notion that self is a process suggests that it can come and go, for instance, when the relevant processes are suspended during sleep,[8] perhaps even with the switch from an inward to an outward focus of attention. This idea of the self as a transient thing has also been put forward by the philosopher Galen Strawson, who has proposed "that many mental selves exist, one at a time and one after another, like pearls on a string" ([60], p. 424). This poetic metaphor asserts a number of things. First, that the self is not continuous, immutable, and immortal (as Descartes and many others have imagined, and as Hume and others have questioned), and second, that "selves" are nevertheless "things" worthy of study, and perhaps capable of emulation.

What we particularly like about Strawson's approach is that he provides some helpful suggestions as to how we might proceed with the study of self, highlighting five questions ([60], p. 406):

1. *The phenomenological question—what is the nature of the sense of the self?*
2. *The local phenomenological question—what is the nature of the human sense of the self?*
3. *The general phenomenological question—are there other possibilities, when it comes to a sense of the self, e.g., can we describe the minimal case?*
4. *The conditions question—what are the grounds or preconditions of possession of a sense of the self?*

---

demonstrating that "my body is wherever there is something to be done" (Merleau-Ponty, [36] p. 291) and providing new ways to test hypotheses about the self.

8 This was proposed by David Hume [28], for whom, if the stream of perceptions is turned off, as happens in sleep, the self ceases to exist, and by John Locke [33], for whom self was a manifestation of consciousness, which, in turn, requires an awake mind. Some elements of Locke's view of self, which saw identity as arising from learning and memory, are close to the ideas of the extended and narrative selves discussed in this chapter.

5. *The factual (metaphysical) question—is there (could there be) such a thing as the self?*

Questions 1 and 2 are psychological in nature, and we think that we can make progress on these through empirical exploration[9] of the facets of self and their variability across the population, taking into account, in particular, developmental and neurological differences. Indeed, a wealth of literature already exists on these topics going back to the earliest days of psychological investigation, some of which is discussed in brief below.

Question 3 might direct us to the panoply of animal life as an interesting place to look for the presence of other kinds of self (and pending the discovery of any extraterrestrial selves). Comparative cognition offers many interesting insights, as well as proposals for how we might test for similar facets of self across species. However, with robotics, we also have the possibility of building new kinds of self, including candidate minimal selves, for which we might adopt some of the cross-species yardsticks identified by comparative studies.

Question 4 speaks to another kind of enquiry, namely as to whether there are any necessary conditions restricting the possibility of an entity possessing a self. One requirement we might posit is a body-world boundary and the ability to sense and maintain the internal milieu, while another might be the possession of a particular kind of cognitive architecture in which there are processes that have the capacity to monitor and predict other internal processes. These ideas will be discussed further below.

Finally, question 5 seems to be largely philosophical, however, we think that progress could also be made via a synthetic approach. Specifically, once we have built a robot that exhibits some relevant phenomena of self, we can ask whether a particular conception of self, for instance, Strawson's string of pearls, is useful or not.[10] In-

---

[9] We should admit here that Strawson intends the more restricted philosophical sense of phenomenology as a form of systematic reflection on the structure of experience. We prefer to interpret the challenge of describing the nature of self from a more empirical perspective as phenomena associated with self that could be accessible to methods in psychology and cognitive neuroscience.

[10] Note that, for a theory or concept of self to be useful, we would not consider that the self has to be emergent in a strong sense (that is, not reducible to lower level phenomena), but rather it has to serve a useful explanatory function in our psychological theory. In other words, the con-

deed, we will have an instantiation of a specific theory of self *as a machine*, whose inner workings will be far more accessible than those of a human mind (see [38]). Such a robot should provide an insightful tool for advancing both the philosophical and scientific understanding of the phenomenon of self-hood.

As we peruse Strawson's questions, we think it becomes evident that synthetic psychology could have a lot to say. For instance, on the question of the constitutive conditions, we can build synthetic systems that match the proposed requirements, then apply our phenomenological and Garland tests: Does it behave as though it has a self? Do others see it as having a self? We can also make progress on this question of the minimal form of the target phenomenon—what is the simplest robot that could qualify for self-hood? Let's build it and study it. On the issue of architecture, we can seek to identify a decomposition of the systems underlying the human self that, when suitably replicated in a robot, gives rise to self-like phenomena; this seems to us to be a tractable, if ambitious, challenge.

Note that if selves are transient, as Strawson and others have proposed, we do still need to explain why the experience of self is one of continuity—that you feel you are the same self yesterday, today, tomorrow. Here, we can appeal to the continuity of the body (and the localization of the self within the body) as providing much of the necessary continuity. We can also look to episodic memory and imagination as allowing the instantaneous self to roam in time, recollecting itself as it once was and imagining itself as it might yet be, thus creating an experience of self that can step outside the present and conceive of itself as enduring. Finally, we can consider semantic memory and narrative as providing the basis for a stable self-concept (beliefs and stories about the self). These ideas can also be investigated in our robotic models.

---

cept of self as explicated and realized in machine form should help us to provide useful accounts of human (or machine) cognition and behavior. See Verschure and Prescott [72] for a discussion of theory building and the role of synthetic approaches in the sciences of mind and brain.

4 A "systems" view of self

The plan to create a synthetic robot self becomes more plausible if we can find good evidence for a "systems" view of self in psychology and cognitive neuroscience. If this human "self-system" is at least weakly modular,[11] then we can proceed by building the necessary components, then integrating them with each other and within our robot control architecture, gradually approaching a model of the complete self.

The psychological literature related to the self is vast, and we will not seek to summarize it here. One starting point is the often cited proposal made by the cognitive psychologist Ulrich Neisser [40, 41], who suggested five different kinds of self-knowledge:

"The *ecological self* is the individual situated in and acting upon the immediate physical environment. […]. The *interpersonal self* is the individual, engaged in social interaction with another person. […]. The *conceptual self*, or self-concept, is a person's mental representation of his/her own (more or less permanent) characteristics. […] The *temporally extended self* is the individual's own life-story as he/she knows it, remembers it, tells it, projects it into the future. The *private self* appears when the child comes to understand and value the privacy of conscious experience […]" ([41], pp. 18-19, our italics). Table 1 builds on Neisser's five-way split, conceiving of each of these as a sub-system of the self and relating each to some psychological phenomena that can provide benchmarks for the existence of that aspect of self in a person or robot. We have also followed Shaun Gallagher [20], Marc Jeannerod [30] and others by adding agency—the *agential self*. The systems view asserts that some sense of self can emerge in the absence of some of these components and that some aspects of self, perhaps particularly the private self, could emerge from the interaction of these components without being explicitly designed, i.e., the sum is more than its parts.

---

[11] Modularity is itself a topic that is widely debated within the cognitive sciences. Again, we consider that the synthetic approach can help answer some of the longstanding questions about how distributed vs. modular human minds/brains are. Our view is that the distributed nature of the brain can be over-stated. The brain is a layered architecture [49], and as such, there *is* significant replication of function and some redundancy across these layers, however, there is also localization of function and specific local or repeated circuits that perform roles that can be clearly described and differentiated.

Phenomena of self	**Component of self**
Sensing the body Distinguishing yourself from the world Having a point-of-view Actively seeking sensory information	Ecological
Having emotions, drives and motivations Selecting actions that generate integrated behavior Knowing what events you have caused in the world	Agential
Having awareness of where you are Having awareness of a personal past and future Self-recognition (e.g., in a mirror) Knowing what you will do next	Extended
Learning by imitation Sharing attention Seeing others as selves Imagine other points-of-view	Interpersonal
Having beliefs about who you are (a self-concept) Having personal goals Having a life story (a narrative)	Conceptual
Having experience *Having a feeling of being something* *Having a unitary stream of consciousness* *Having a sense of choice* *Having a feeling of being the same thing over time*	*Private*

Table 1 Some of the phenomena of self and how these might be grouped into different self-components based on Neisser [40,41], Gallagher [20] and others. These sub-systems are assumed to be weakly modular but with significant interdependencies. The private self is in italics since it reflects first-person phenomena that may be emergent properties of the wider system. This decomposition is intended as a hypothesis to be investigated, refuted and revised using both analytical (empirical) and synthetic approaches.

5 A diversity of selves across the life-span, the population, and the animal kingdom

There is evidence to support this "systems" view of self from developmental psychology, neuroscience, and comparative psychology, which we will briefly review next.

From the study of human development, it is clear that very young infants have a sense of their ecological selves, for example, having a self-other distinction. This may emerge through exploration of the body in the womb. The fetus explores and discovers its body through "motor babbling"; it also touches itself, and the experience of skin-on-own-skin, or "double touch", is different from the experience of touching parts of the mother [52]. These activities allow the unborn child to learn the extension and limits of its own body. The emerging ability to control its own body, and to distinguish when a sensory event was caused by its own action, can also provide the newborn with some pre-reflective sense of agency (along the lines proposed by Jeannerod [30]). Agency in older children is often studied in the context of executive function and self-regulation, for example, the ability to withhold actions, show cognitive flexibility, or control emotional expression; these aspects of agency show multiple phases of development through infancy and the pre-school years [7, 73]. Infantile amnesia, which lasts until we are around two years of age [31], implies that the infant lives in the here and now, lacking a strong sense of its extension in time. The mirror test—recognizing that it is you in a mirror, not another child—is another milestone for the two-year-old [2, 4] that may indicate the beginnings of a reflective self-model. The newborn is a social creature, adapted to bond rapidly with its caregivers, yet significant changes occur in its capacity for sociality in the first year, including the emergence of shared attention, social referencing (looking to adults to understand the

meanings of events), imitation, and wariness of strangers [42]. It is not until a child is around three years of age that it has "theory of mind"—the ability to conceive of another's point-of-view as different from its own [16]. The emergence of this interpersonal self, which is able to interpret the actions and intentions of others, likely builds on capacities of the ecological self to represent and reason about the child's own body. Finally, the conceptual self may emerge from the extended self, through consolidation of episodic memories into semantics—knowledge of the self and the world—and with help from the growing capacity to manipulate concepts and summarize events using language. Prior to the school years, children struggle to assemble coherent descriptions of past episodes [6], but as we grow older, we get more practiced at translating life events into story form, with the most important ones being rehearsed and consolidated to become stable chapters in the emerging self-narrative.

In the neurosciences, there is evidence from the study of neurodiversity and brain damage that also supports the decomposition of the self into component parts. Many conditions can impact on the sense of the ecological self: a disturbed body model can generate sensory neglect [68], or the sense that a part of your body does not belong to you (see [10]). Disorders of the hypothalamus, the basal ganglia, limbic system and prefrontal cortex can disrupt motivation, action integration and the experience of agency [29, 30, 43]. Damage to areas such as the temporal lobe, particularly the hippocampal system, can cause loss of the sense of place, or of the ability to think about the past or future, whilst sparing the core sense of the self in the here and now [65].[12] Activity in the "default mode" network of cortical sub-systems is also recognized as a critical substrate for the human capacity for "mental time travel" [56]. A well-known example of an altered social self occurs in people with autism, a condition that par-

12 Endel Tulving's patient N.N. exemplifies this point [65]. A traffic accident caused N.N. to experience profound retrograde and anterograde amnesia, nevertheless he could still talk about himself, his experience, his preferences, and so on; he had intact short-term memory and could describe time and events in general terms. He could talk about consciousness, which he described as "being aware of who we are and what we are and where we are" ([65], p. 4). When asked to imagine what he might do tomorrow, however, his mind drew a blank, which he described as being "like swimming in the middle of a lake. There's nothing there to do hold you up or do anything with" ([65], p. 4). Like other patients with amnesia, N.N. could be described as "marooned in the present" [32] or as having a self that has lost much of its "temporal thickness" [19].

ticularly impacts on the ability to understand others as social actors [5], whilst leaving intact other aspects of self (however, see [67]). The phenomenon of multiple personality disorder (e.g., [58]) shows the possibility that the self can assemble itself into one identity at one time, and into a very different one a few minutes later, with no shared consciousness or memory. This speaks to the constructed nature of the self and to its dynamical character as well. Specifically, if we think of identity as a stable attractor for the self system, then, in the unusual case of multiple personalities, the system is bi- or multi-stable and able to flip between different internally coherent, but mutually inconsistent, conceptions of self.

Comparative psychology also demonstrates variety in the nature of self (if we accept that animals can have selves). A self-other distinction, along with an ability to recognize the consequences of your actions, and hence some form of minimal self, may be shared by all bilateral multi-celled animals (see below). On the other hand, the capacity to conceive of the self as extending into the future and the past is far less universal and may only be well-developed in a limited number of animal groups, including some of the larger-brained mammals and birds [61]. The ability to voluntarily search in autobiographical memory for traces of particular events may be specific to humans having evolved in early homo lineages [17]. Evidence of a reflective self-model, as demonstrated by the mirror test, has also been shown in only a limited number of species, including great apes, dolphins, orca whales, elephants, and one species of bird (Eurasian magpies) [51]. The presence of an interpersonal self that has theory of mind, which has been extensively investigated only in primates, may also be confined to animals that have an expanded neocortex [64].

6 A minimal robotic self?

As noted earlier, one of the questions we would like to address through the synthetic approach concerns the possibility of a minimal self. Shaun Gallagher [20] reviews a number of proposals for minimal selves, identifying two key aspects, body ownership and agency, similar to the ecological and agential sub-systems noted in Table 1.

He suggests, following Bermúdez [8], that the sense of self can be non-conceptual, pre-reflective, confined to the present, and a transient entity like one of Strawson's pearls.

Jan Tani [62] has sought to create such a transient self for a mobile robot through a simple layered control system consisting of a perception module, an association module, and a prediction module. The robot was tasked with following a wall whilst searching for colored landmarks; the actions of the robot consist of steering by controlling left and right wheel-speeds and choosing whether to allocate visual attention to wall-following or to landmark searching. The robot monitors the reliability of its own predictions and uses this to arbitrate between control by the "bottom-up" sensory module and that by the "top-down" prediction module. Tani proposes that a form of self emerges when the predictions of the top-down module diverge from those of the sensory module, resulting in a period of dynamic instability, and that this "self" disappears when the prediction and sensory modules transition to a period of coherence.

Tani draws analogies to mammalian brain systems, however, the simple control system that he describes could be compared to much simpler nervous systems, for example, the nerve nets of some jellyfish can be conceived of as forming layered architectures in which distinct distributed networks compete for control of the motor system [50]. The earliest bilaterian animals, whose existence in the Precambrian era more than 540 million years ago is evidenced by fossils of their foraging trails, likely possessed internal organs, tentacle-like appendages, multiple sensors, and a nervous system that included a central ganglion, sometimes referred to as the "archaic brain" (see [50] for review). Modern day worms, including animals as simple as *C. Elegans*, have shown associative learning and the ability to use sensory signals to predict aversive chemicals and the presence of absence of food [3]. If monitoring the divergence between internal expectations about the world and sensory experience can give rise to a self, then perhaps minimal selves were present in some of the first mobile multicellular animals.

Tani's model is based on the hypothesis that the self requires a process that has an internal state that can evolve according to its own dynamics without being too tightly coupled to the world—the predictions of the system can drift from accurately forecasting the

world, and at this point, the robot obtains a self. However, all animals with nervous systems interoceptively sense their bodies at the same time as they exteroceptively sense the environment; the patterns of sensory signals from the internal milieu, which will have very different dynamics from those of the sensed external world, thus already provide a basis for pre-reflectively distinguishing self from other.

7 A biomimetic cognitive architecture for the robot self

In Sheffield, we have been building and testing brain-based robots, as experiments in synthetic psychology, since the mid-1990s, devising a number of models of brain architecture based on principles of layered control [49] and inspired by neurobehavioral studies of active sensing in rodents [48]. For the past seven years, together with European colleagues, we have also been incorporating models of key brain systems into a brain-inspired control architecture for the iCub robot (Figure 1) called *distributed adaptive control* (DAC), developed by Paul Verschure and colleagues [69-71].

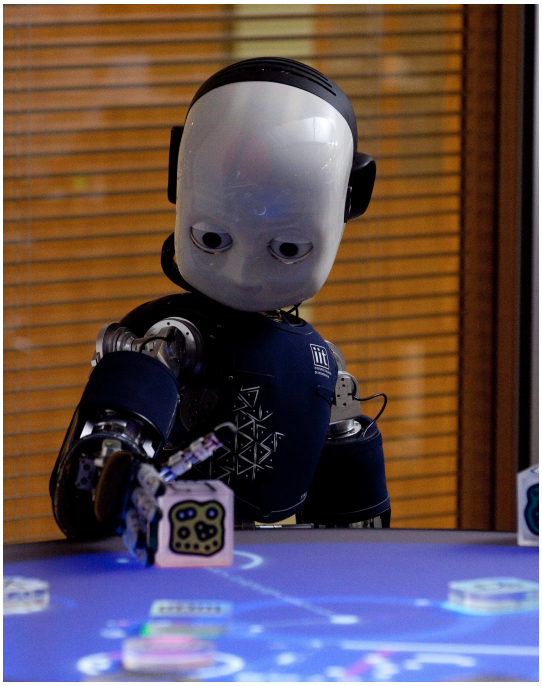


Fig. 1 The iCub humanoid robot. A biomimetic robot platform for embodied testing of theories of general human intelligence developed by European researchers led by the Italian Institute of Technology. Picture from Sheffield Robotics.

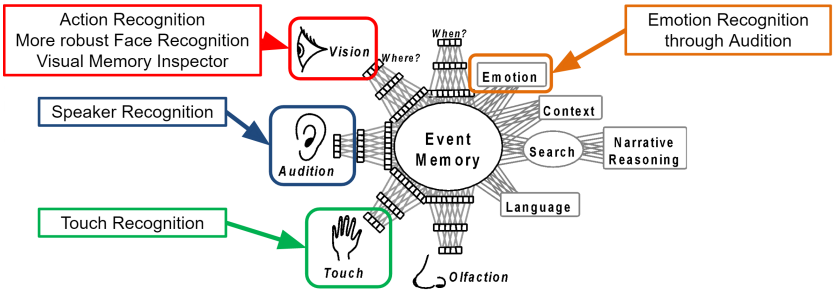
DAC is a high-level conceptual scheme that seeks to capture the cognitive architecture of the human brain and consists of four tightly coupled layers: *soma, reactive, adaptive* and *contextual*. Across these layers, there are three functional columns of organization: The first comprises the sensory, perceptual and memory sub-systems relating to the *world*, the second the interoceptive, motivational and memory sub-systems related to the *self*, and the third sub-systems that operate on the world through *action*. These DAC sub-systems do not directly map on to specific neural substrates, however, significant progress has been made relating parts of the DAC architecture to different brain sub-systems and circuits [47, 69]. Recent efforts to create a multi-faceted robot sense of self for iCub, using DAC, are detailed in [39]; here, we briefly summarize the architecture and some of the self-related capabilities it enables.

In DAC, the *somatic* layer corresponds to the body and provides access to exteroceptive, interoceptive, and proprioceptive signals from, respectively, the environment, internal processes and regulatory systems, and the motor/effector system. The *reactive* layer instantiates multiple fast, reflexive sensorimotor loops that support behaviors linked to needs; these loops are stability-seeking processes that reduce drives through action. The *adaptive* layer extends the sensorimotor loops of the reactive layer to make use of learned contingencies and to allow actions to be associated with states of the world. The adaptive layer is thus part of the solution to the symbol grounding problem, through the acquisition of mappings from internal states to world states. Whereas the adaptive layer operates largely in the here and now, the *contextual* layer adds the ability to store and retrieve short- and long-term memories, linked to goal achievement, that can act as action plans to be triggered by sensory contexts and that can be chained to create behavior sequences. This layer also includes predictive systems that can forecast the future state of the world based on action plans. Contextual layer systems can also encode and retrieve event memories and form abstract representations of events in narrative form that allow the robot to summarize and communicate about past episodes.

The DAC architecture generates aspects of the *ecological self* through interoceptive processes that maintain a model of the robot's physical parts and the geometry of its current body pose, and exteroceptive processes that monitor the robot's immediate surroundings. For example, using somatotopic maps modelled on human primary sensory cortex, and techniques such as self touch, Giorgio Metta, Matej Hoffmann and colleagues have developed methods that allow the iCub to learn its own body model [26], and recalibrate its knowledge of its own geometry [54]. Additionally, by combining vision with tactile sensing and with proprioception, iCub is able to develop a sense of peripersonal space that allows it to predict contacts with objects before they happen [53]. This foundation provides the beginnings of an ecological self that can be used to distinguish self from other, plan safe movement trajectories, and reason about the capacity for movement of others (see more below).

In Sheffield, we have been working to develop an episodic or *event* memory system for the DAC adaptive and contextual layers that can

contribute to a robotic *extended self*. Our hypothesis is that event memory can be usefully considered as an attractor network operating in a latent (hidden) variable space whose dimensions encode salient characteristics of the physical and social world in a highly compressed fashion [18]. According to this view, the operation of perceptual systems in the adaptive and contextual layers can be analogized to learning processes that identify psychologically meaningful latent variable descriptions. Instantaneous memories then correspond to points in this latent variable space and event memories to trajectories through this space. A single latent feature space can be used to represent memories across multiple sensory modalities thus providing sensory fusion. This enhances compression as coupled signals among heterogenous modalities are discovered and represented in a common set of latent variables. This can also be thought of as concept discovery—the identification of underlying invariance in patterns of multi-modal sensory flow. The current implementation, illustrated in Figure 2, demonstrates effective memory formation and retrieval of human faces, actions, voices and emotions [12, 14, 35]. Due to its generative nature, and ability to interpolate, the system can also generate fantasy memories from parts of the latent variable space that have not been populated by real data. This leads to the possibility of imagining future events [14]. The ability of the system to reconstruct the sensory pattern associated with a recalled memory [11], retrieved using a verbal cue, suggests that event memory can contribute to the grounding of linguistic symbols in sensorimotor experience.



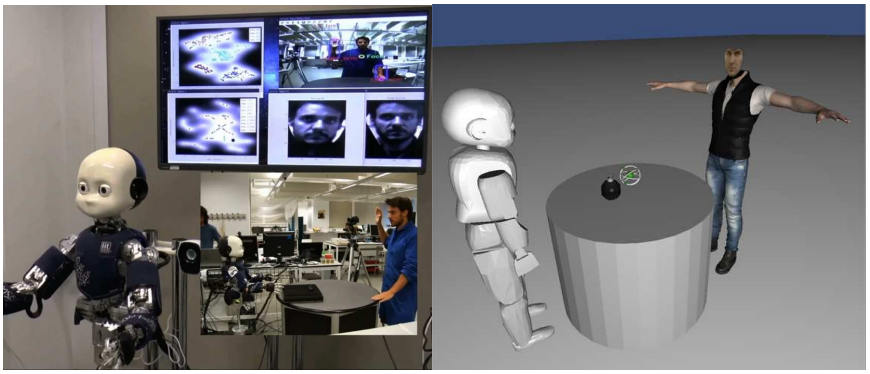


Fig. 2 The synthetic event memory system developed at Sheffield. Our model of event memory integrates across multiple modalities to encode memories as patterns in a low-dimensional latent variable space that can be used to reconstruct past experiences based on partial cues or explicit search. *Top:* The proposed architecture for a synthetic episodic/event memory system based on Rubin [55] and Evans et al. [18]. The highlighted areas show the components that have been constructed to date, which include sub-systems for action, touch, and emotion recognition, for speaker recognition in the visual and auditory modalities, and the ability to display visual memories (the visual memory inspector). *Bottom-left*: iCub operating in real-time to recognize actions and faces. The TV monitor behind the robot shows two latent variable spaces, the visual pre-processing of the camera scene, and the reconstruction of the remembered face based on the recovered memory. *Bottom-right*: a screen-shot from the visual memory inspector, which allows researchers to see iCub's simulation of itself and its perceptual world. Here, iCub represents a face and two objects on the circular table.

Neuroscience research suggests that an effective approach to building the *interpersonal self* could be to use the robot's own internal body models—the ones that underlie the ecological self—to simulate the pose and actions of others. With iCub, we have developed DAC processes that allow the robot to represent the state of the world from a different point-of-view (see [39]), allowing iCub to reason about what a human partner can see and helping the robot to resolve perceptual ambiguities and improve communication. One important human ability that benefits from the interpersonal self is the capacity to learn by imitation. Yiannis Demiris and co-workers have demonstrated that you can build up from motor babbling to a hierarchical learning system that uses forward models, inspired by studies of the primate "mirror neuron" system, to learn by imitation [15]. This system has been used with the iCub to allow it to rapidly

acquire new hand gestures and sequences of actions involved in playing games or solving puzzles.

As shown in Figure 2, a key part of the broader system in which our synthetic event memory operates is the component related to narrative reasoning, this is one of the sub-systems that generate the *conceptual self*. Peter Dominey and colleagues (see, e.g., [4]) have been working to model autobiographical memory and narrative construction using an acquired grammar, together with compact and structured representations of iCub's interaction history. Using this narrative system, iCub can recall and discuss past events, including some of its past interactions with people, from a first-person perspective. One longer-term goal is to integrate this narrative construction process with the event memory system developed in Sheffield such that linguistic descriptions can be abstracted from representations of events as attractor patterns in latent variable space. Using the generative capabilities of the event memory, narratives could also be played out, and "grounded", via reconstruction as simulated sensory scenes.

In sum, we have made a start in instantiating some of the different aspects of the sense of self in our iCub robot. The lower layers of the DAC architecture integrate internal and external sensory signals so as to regulate self-correcting control loops based on drives. These sub-systems meet many of the criteria for a minimal self. The upper layers encode representations of past events that can be used to reason about the future and about social others, creating some of the elements that we are seeking for the extended and interpersonal selves. Finally, the narrative system provides the seeds for a self-concept and life story. We have not sought to build a *private self* directly, rather, the plan is to create the rest of the architecture and then see if an impression of the experiential self can emerge from within in our version of the Garland test. Indeed, on a good day, when all of the sub-systems are working properly, interacting with the iCub can begin to feel as though "someone is home", even for the people who have helped to develop the robot's control systems and understand how they operate. On the other hand, it also feels as if we have only just set out on the journey of deconstructing the human self and recreating it in a machine. Indeed, as Turing wrote at the end of *Computing Machinery and Intelligence*— "we can only see a short dis-

tance ahead, but we can see plenty there that needs to be done" ([66], p. 460).

8 Conclusion

This chapter has argued that the human self is brought into being by the activity of a set of self-processes instantiated by the brain and body and has proposed that we can synthesize an artificial self by developing equivalent sub-systems within an integrated biomimetic cognitive architecture for a humanoid robot. While the various self-processes may be transient, the continuity provided by a physical body, in a human or robot, can provide the basis for the experience of a continued self. This suggests a key role for embodiment, first in establishing a boundary between the self and the world, and second in providing a predictable and consistent setting in which the self awakens to find itself. Beyond this, an extended self, generated by the capacity to remember and imagine, allows the self to escape from the island of the present, while abstraction and narrative allow it to construct and maintain a coherent set of beliefs and stories about itself. To evaluate the possibility of a robot self, we have suggested a version of the Turing test, extended to include physical embodiment and human-like cognitive architecture, that asks whether people who encounter a robot with synthetic self-processes consider that they have met an entity with a self.

We began the chapter by motivating this work in the context of the criteria for recognizing other minds, and the challenge of benchmarking artificial general intelligence against human. We have concluded by summarizing some initial efforts to create a sense of self for the iCub humanoid robot that has ecological, temporally-extended, interpersonal and narrative components set within a multi-layered model of mind.

Acknowledgements

The preparation of this chapter was supported by funding from the EU Seventh Framework Programme as part of the projects *Experimental Functional Android Assistant* (EFAA, FP7-ICT-270490) and *What You Say Is What You Did* (WYSIWYD, FP7-ICT-612139) and the EU H2020 Programme as part of the *Human Brain Project* (HBP-SGA1, 720270). We are particularly grateful to Paul Verschure, Peter Dominey, Giorgio Metta, Yiannis Demiris and the other members of the WYSIWYD and EFAA consortia, and to our colleagues at the University of Sheffield who have helped us to develop memory systems for the iCub, particularly Uriel Martinez, Andreas Damianou, Neil Lawrence, Luke Boorman and Matthew Evans. The Sheffield iCub was purchased with the support of the UK Engineering and Physical Science Research Council (EPSRC).

References

[1] Abel, C. (2014). The Extended Self: Architecture, Memes and Minds. Manchester: Manchester University Press.

[2] Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology, 5*(4), 297-305.

[3] Ardiel, E. L., & Rankin, C. H. (2010). An elegant mind: Learning and memory in *Caenorhabditis elegans*. *Learning & Memory, 17*(4), 191-201. doi:10.1101/lm.960510

[4] Bard, K. A., Todd, B. K., Bernier, C., Love, J., & Leavens, D. A. (2006). Self-awareness in human and chimpanzee infants: What Is measured and what Is meant by the mark and mirror test? *Infancy, 9*(2), 191-219. doi:10.1207/s15327078in0902_6

[5] Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition, 21*, 37-48.

[6] Bauer, P. J. (2012). The life I once remembered: the waxing and waning of early memories. In D. Berntsen & D. C. Rubin (Eds.), *Understanding Autobiographical Memory* (pp. 205-225). Cambridge: CUP.

[7] Bell, M. A., & Deater-Deckard, K. (2007). Biological Systems and the Development of Self-Regulation: Integrating Behavior, Genetics, and Psychophysiology. *Journal of Developmental & Behavioral Pediatrics, 28*(5).

[8] Bermúdez, J. (1988). *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

[9] Blakemore, S. (2003). Consciousness in meme machines. *Journal of Consciousness Studies, 10*(4-5), 19-30.

[10] Blakeslee, S., & Blakeslee, M. (2007). *The Body has a Mind of its Own*. New York: Random House.

[11] Camilleri, D., Damianou, A., Jackson, H., Lawrence, N., & Prescott, T. J. (2016). iCub visual memory inspector: Visualising the iCub's thoughts. In N. F. Lepora, A. Mura, M. Mangan, P. F. M. J. Verschure, M. Desmulliez, & T. J. Prescott (Eds.), *Biomimetic and Biohybrid Systems, the 5th International Conference on Living Machines* (pp. 48-57). Berlin: Springer LNAI.

[12] Camilleri, D., & Prescott, T. J. (2017). *Action recognition with unsynchronised multi-sensory data.* Paper presented at the 7th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB) 2017, Lisbon, Portugal.

[13] Cangelosi, A., Schlesinger, M., & Smith, L. B. (2015). *Developmental robotics: From babies to robots*. Cambridge, MA: MIT Press.

[14] Damianou, A., Henrik, C., Boorman, L., Lawrence, N. D., & Prescott, T. J. (2015). A top-down approach for a synthetic autobiographical memory system. In S. Wilson, T. J. Prescott, A. Mura, & P. F. M. J. Verschure (Eds.), *Biomimetic and Biohybrid Systems, the 4th International Conference on Living Machines* (Vol. 9222, pp. 280-292). Berlin: Springer LNAI.

[15] Demiris, Y., Aziz-Zadeh, l., & Bonaiuto, J. (2014). Information Processing in the Mirror Neuron System in Primates and Machines. *Neuroinformatics, 12*(1), 63–91.

[16] Doherty, M. (2009). *Theory of Mind: How Children Understand Others' Thoughts and Feelings*. Hove: Psychology Press.

[17] Donald, M. (2012). Evolutionary origins of autobiographical memory: A retrieval hypothesis. In D. Berntsen & D. C. Rubin (Eds.), *Understanding Autobiographical Memory* (pp. 269-289). Cambridge: CUP.

[18] Evans, M. H., Fox, C. W., & Prescott, T. J. (2014). Machines Learning - Towards a New Synthetic Autobiographical Memory. In A. Duff, N. Lepora, A. Mura, T. Prescott, & P. M. J. Verschure (Eds.), *Biomimetic and Biohybrid Systems, the 3rd International Conference on Living Machines* (Vol. 8608, pp. 84-96). Berlin: Springer LNAI.

[19] Friston, K. (2017). The mathematics of mind-time. *Aeon*.

[20] Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4*(1), 14–21.

[21] Gardner, H. (2006). *Multiple Intelligences: New Horizons*. New York: Basic Books.

[22] Goertzel, B., & Pennachin, C. (2007). *Artificial General Intelligence*. New York: Springer.

[23] Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines, 1*, 43–54.

[24] Harnad, S. (1994). Does the mind piggy-back on robotic and symbolic capacity? In H. L. Morowitz & J. L. Singer (Eds.), *The Mind, the Brain, and Complex Adaptive Systems, Santa Fe Institute Studies in Complexity XXII* (pp. 204-220). Boston: Addison Wesley.

[25] Hauser, L. (1993). Reaping the worldwind: Reply to Harnad's "Other bodies, other minds". *Minds and Machines, 3*, 219–238.

[26] Hoffmann, M., Straka, Z., Farkas, I., Vavrecka, M., & Metta, G. (2017). Robotic homunculus: Learning of artificial skin representation in a humanoid robot motivated by primary somatosensory cortex. *IEEE Transactions on Cognitive and Developmental Systems, PP*(99), 1-1. doi:10.1109/TCDS.2017.2649225

[27] Hood, B. (2012). *The Self Illusion: Why There is No 'You' Inside Your Head*. London: Constable and Robinson.

[28] Hume, D. (1740). A Treatise on Human Nature.

[29] Humphries, M. D., & Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress In Neurobiology, 90*(4), 385-417. doi:10.1016/j.pneurobio.2009.11.003

[30] Jeannerod, M. (2003). The mechanism of self-recognition in humans. *Behavioural Brain Research, 142*(1), 1-15. doi:doi.org/10.1016/S0166-4328(02)00384-4

[31] Lambert, F. R., Lavenex, P., & Lavenex, P. B. (2017). The "when" and the "where" of single-trial allocentric spatial memory performance in young children: Insights into the development of episodic memory. *Developmental Psychobiology, 59*(2), 185-196. doi:10.1002/dev.21479

[32] Lidz, T. (1942). The amnesic syndrome. *Archives of Neurology and Psychiatry, 47*, 588-605.

[33] Locke, J. (1777). An Enquiry Concerning Human Understanding.

[34] Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science, 15*(4), 151-190. doi:10.1080/09540090310001655110

[35] Martinez-Hernandez, U., Damianou, A., Camilleri, D., Boorman, L. W., Lawrence, N., & Prescott, T. J. (2016, 3-7 Dec. 2016). *An integrated probabilistic framework for robot perception, learning and memory.* Paper presented at the 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO), Qingdao, China. pp. 1796-1801

[36] Merleau-Ponty, M. (1945/1962). *Phénoménologie de la Perception* (C. Smith, Trans.). London: Routledge.

[37] Metzinger, T. (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

[38] Mitchinson, B., Pearson, M., Pipe, T., & Prescott, T. J. (2011). Biomimetic robots as scientific models: A view from the whisker tip. In J. Krichmar & H. Wagatsuma (Eds.), *Neuromorphic and Brain-based Robots* (pp. 23-57). Boston, MA: MIT Press.

[39] Moulin-Frier, C., Fischer, T., Petit, M., Pointeau, G., Puigbo, J. Y., Pattacini, U., Low, S. C., Camilleri, D., Nguyen, P., Hoffmann, M., Chang, H. J., Zambelli, M., Mealier, A. L., Damianou, A., Metta, G., Prescott, T. J., Demiris, Y.,

Dominey, P. F., & Verschure, P. F. M. J. (2017). DAC-h3: A Proactive Robot Cognitive Architecture to Acquire and Express Knowledge About the World and the Self. *IEEE Transactions on Cognitive and Developmental Systems, PP*(99), 1-1. doi:10.1109/TCDS.2017.2754143

[40] Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology, 1*, 35-59. doi:10.1080/09515088808572924

[41] Neisser, U. (1995). Criteria for an ecological self. In P. Rochat (Ed.), *The Self in Infancy: Theory and Research*. Amsterdam: Elsevier.

[42] Nelson, K. (2007). *Young Minds in Social Worlds: Experience, Meaning and Memory*. Cambridge, MA: Harvard University Press.

[43] Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford: OUP.

[44] Pointeau, G., & Dominey, P. F. (2017). The Role of Autobiographical Memory in the Development of a Robot Self. *Frontiers in Neurorobotics, 11*, 27.

[45] Prescott, T. J. (2015, 21st March). Me in the Machine. *New Scientist,* 36-39.

[46] Prescott, T. J., Ayers, J., Grasso, F. W., & Verschure, P. F. M. J. (2016). Embodied models and neurorobotics. In M. A. Arbib & J. J. Bonaiuto (Eds.), *From Neuron to Cognition via Computational Neuroscience* (pp. 483-512). Cambridge, MA: MIT Press.

[47] Prescott, T. J., Lepora, N., & Verschure, P. F. M. J. (2018). *The Handbook of Living Machines: Research in Biomimetic and Biohybrid Systems*. Oxford, UK: OUP.

[48] Prescott, T. J., Mitchinson, B., Lepora, N. F., Wilson, S. P., Anderson, S. R., Porrill, J., Dean, P., Fox, C. W., Pearson, M. J., Sullivan, J. C., & Pipe, A. G. (2015). The robot vibrissal system: Understanding mammalian sensorimotor co-ordination through biomimetics. In P. Krieger & A. Groh (Eds.), *Sensorimotor Integration in the Whisker System* (pp. 213-240): Springer New York.

[49] Prescott, T. J., Redgrave, P., & Gurney, K. N. (1999). Layered control architectures in robots and vertebrates. *Adaptive Behavior, 7*(1), 99-127.

[50] Prescott, T. J. (2007). Forced moves or good tricks in design space? Landmarks in the evolution of neural mechanisms for action selection. Adaptive Behavior, 15(1), 9-31.
[51] Prior, H., Schwarz, A., & Gunturkun, O. (2008). Mirror-Induced Behavior in the Magpie (Pica pica): Evidence of Self-Recognition. *PLoS Biol, 6*(8), e202.
[52] Rochat, P. (2001). *The Infant's World*. Cambridge, MA: Harvard University Press.
[53] Roncone, A., Hoffmann, M., Pattacini, U., Fadiga, L., & Metta, G. (2016). Peripersonal Space and Margin of Safety around the Body: Learning Visuo-Tactile Associations in a Humanoid Robot with Artificial Skin. *PLoS ONE, 11*(10), e0163713. doi:10.1371/journal.pone.0163713
[54] Roncone, A., Hoffmann, M., Pattacini, U., & Metta, G. (2014, May 31 2014-June 7 2014). *Automatic kinematic chain calibration using artificial skin: Self-touch in the iCub humanoid robot.* Paper presented at the 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 2305-2312
[55] Rubin, D. C. (2006). The basic-systems model of episodic memory. *Perspectives on Psychological Science, 1*(4), 277-311.
[56] Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron, 76*(4), 10.1016/j.neuron.2012.1011.1001. doi:10.1016/j.neuron.2012.11.001
[57] Searle, J. (1990). Is the brain's mind a computer program? *Scientific American, 262*(1), 20-25.
[58] Silberman, E. K., Putnam, F. W., Weingartner, H., Braun, B. G., & Post, R. M. (1985). Dissociative states in multiple personality disorder: A quantitative study. *Psychiatry Research, 15*(4), 253-260. doi:10.1016/0165-1781(85)90062-9
[59] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general

reinforcement learning algorithm. *arXiv:1712.01815v1 [cs.AI] 5 Dec 2017*.

[60] Strawson, G. (1997). The self. *Journal of Consciousness Studies, 4*(5/6), 405-428.

[61] Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences, 30*(3), 299-313. doi:10.1017/S0140525X07001975

[62] Tani, J. (1998). An interpretation of the 'self' from the dynamical systems perspective: a constructivist approach. *Journal of Consciousness Studies, 5*, 516–542.

[63] Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(1), 42. doi:10.1186/1471-2202-5-42

[64] Towner, S. (2010). Concept of mind in non-human primates. *Bioscience Horizons: The International Journal of Student Research, 3*(1), 96-104. doi:10.1093/biohorizons/hzq011

[65] Tulving, E. (1985). Memory and consciousness. *Canadian Journal of Psychology, 26*(1), 1-12.

[66] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59(236)*, 433-460.

[67] Uddin, L. Q. (2011). The self in autism: An emerging view from neuroimaging. *Neurocase, 17*(3), 201-208. doi:10.1080/13554794.2010.509320

[68] Vallar, G. (1998). Spatial hemineglect in humans. *Trends in Cognitive Sciences, 2*(3), 87-97. doi:10.1016/S1364-6613(98)01145-0

[69] Verschure, P. F. M. J. (2012). Distributed Adaptive Control: A theory of the Mind, Brain, Body Nexus. *Biologically Inspired Cognitive Architectures, 1*, 55-72. doi:10.1016/j.bica.2012.04.005

[70] Verschure, P. F. M. J., Krose, B., & Pfeifer, R. (1992). Distributed Adaptive Control: The self-organization of structured behavior. *Robotics And Autonomous Systems, 9*, 181--196.

[71] Verschure, P. F. M. J., Pennartz, C. M. A., & Pezzulo, G. (2014). The why, what, where, when and how of goal-directed choice: neuronal and computational principles.

Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369(1655). doi:10.1098/rstb.2013.0483

[72] Verschure, P. F. M. J., & Prescott, T. J. (2018). A Living Machines approach to the sciences of mind and brain. In T. J. Prescott, N. Lepora, & P. F. M. J. Verschure (Eds.), *The Handbook of Living Machines: Research in Biomimetic and Biohybrid Systems*. Oxford, UK: OUP.

[73] Zelazo, P. D. (2004). The development of conscious control in childhood. *Trends in Cognitive Sciences, 8*(1), 12-17. doi:10.1016/j.tics.2003.11.001

[74] Braitenberg, V. (1986). Vehicles: Experiments in Synthetic Psychology. Cambridge, MA: MIT Press.

[75] Hofstadter, D. (2007). I am a Strange Loop. New York: Basic Books.