PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Cite this article: Verschure PFMJ. 2016 Synthetic consciousness: the distributed adaptive control perspective. *Phil. Trans. R. Soc. B* **371**: 20150448.
http://dx.doi.org/10.1098/rstb.2015.0448

Accepted: 18 May 2016

One contribution of 13 to a theme issue 'The major synthetic evolutionary transitions'.

Subject Areas:
neuroscience

Keywords:
consciousness, distributed adaptive control, neuronal substrate, memory, normativity, social brain

Author for correspondence:
Paul F. M. J. Verschure
e-mail: paul.verschure@upf.edu

Synthetic consciousness: the distributed adaptive control perspective

Paul F. M. J. Verschure[1,2]

[1]Laboratory of Synthetic Perceptive, Emotive and Cognitive Systems, Center of Autonomous Systems and Neurorobotics, Universitat Pompeu Fabra, Barcelona, Spain
[2]ICREA—Institució Catalana de Recerca i Estudis Avançats, 08018 Barcelona, Spain

PFMJV, 0000-0003-3643-9544

Understanding the nature of consciousness is one of the grand outstanding scientific challenges. The fundamental methodological problem is how phenomenal first person experience can be accounted for in a third person verifiable form, while the conceptual challenge is to both define its function and physical realization. The distributed adaptive control theory of consciousness (DACtoc) proposes answers to these three challenges. The methodological challenge is answered relative to the hard problem and DACtoc proposes that it can be addressed using a convergent synthetic methodology using the analysis of synthetic biologically grounded agents, or quale parsing. DACtoc hypothesizes that consciousness in both its primary and secondary forms serves the ability to deal with the hidden states of the world and emerged during the Cambrian period, affording stable multi-agent environments to emerge. The process of consciousness is an autonomous virtualization memory, which serializes and unifies the parallel and subconscious simulations of the hidden states of the world that are largely due to other agents and the self with the objective to extract norms. These norms are in turn projected as value onto the parallel simulation and control systems that are driving action. This functional hypothesis is mapped onto the brainstem, midbrain and the thalamo-cortical and cortico-cortical systems and analysed with respect to our understanding of deficits of consciousness. Subsequently, some of the implications and predictions of DACtoc are outlined, in particular, the prediction that normative bootstrapping of conscious agents is predicated on an intentionality prior. In the view advanced here, human consciousness constitutes the ultimate evolutionary transition by allowing agents to become autonomous with respect to their evolutionary priors leading to a post-biological Anthropocene.

This article is part of the themed issue 'The major synthetic evolutionary transitions'.

1. Introduction

> 'That is very hokey!'
>
> *Anonymous physicist participating in a working group analyzing HED data at ICSB, UC Santa Barbara, 2012 to the question:*
>
> 'Is consciousness not more fundamental than the Higgs boson?'

Understanding the nature of consciousness is the last grand outstanding scientific challenge humans are facing. Since the late Eighties the study of consciousness is again considered scientifically respectable after falling from grace in the early twentieth century in the scientism craze of behaviourism. This resurgence is largely due to the engagement of Francis Crick and Gerald Edelman, who with their added gravitas as Nobel laureates and as direct competitors each advocated opposing strategies to crack the last riddle of biology. The former proposed to identify the Neural Correlate(s) of Consciousness (NCC) [1], while the latter advanced a theory-based approach including the use of synthetic systems or, so-called, brain-based devices [2]. The study of the NCC has given rise to a flurry of research, which unsurprisingly has uncovered a large universe of

THE ROYAL SOCIETY PUBLISHING

correlations between states of mind and brain and consciousness (see [3,4] for recent reviews), while the latter approach has dwindled. Here I will attempt to rectify this imbalance and advance the distributed adaptive control theory of consciousness (DACtoc). Before turning to DACtoc, we have to overcome the interlinked obstacles of the, so-called, hard problem/explanatory gap and its associated panpsychism in order to assure that we actually have something relevant to explain. After that I will provide a framework, which integrates the main views on consciousness, providing a backdrop for DACtoc and its predictions and applications. At the centre of DACtoc stands a synthetic methodology: we will understand biological minds by building them.

The physiologist Emil Du Bois-Reymond observed in 1872: 'What conceivable connection exists between definite movements of definite atoms in my brain on the one hand, and on the other hand, such primordial, indefinable, undeniable facts as these: I feel pain or pleasure; I taste something sweet, or smell a rose, or hear an organ, or see something red, and the certainty that immediately follows: Therefore, I am?' (cited in [5, pp 165–166]. Hippocrates had already proposed that mental states are the result of brain states but how can we explain the experience of these mental states? Descartes answered the scholastics by imposing the efficient causes of the mechanical universe on the study of mind, realizing a reductionist mathematics and physics oriented explanatory framework for *res extensa* at the cost of creating mind–brain dualism, i.e. in his thought experiment doubt is elevated to be the essence of the experiencing mind or the irreducible *res cogitans*. The, so-called, hard problem is an offspring of this dualist tradition. Indeed, today, it is common in the study of consciousness to pay tribute to the division between the Hard Problem and the Easy Problems (HPEP) [6] or the, so-called, explanatory gap [7]: how can we explain the 'raw feel' of the experience of being like something such as Thomas Nagel or his notorious bat [8]? However, behind this profoundness hides a simple syllogism: *Science advances third person descriptions of natural phenomena*; *Conscious experience is first person*, therefore, *Consciousness cannot be described by science*. Hence, the problem is actually one of definition and a number of routes are open to us: declare consciousness a fundamental property of matter and join the current panpsychist movement by converting the explanandum into an axiom, to separate the process of consciousness from its content (quale) or to gain direct third person access to first person experience. DACtoc pursues these latter two options and shows that synthesizing consciousness allows us to describe and explain the first person states of experiencing artificial systems satisfying the Nagel criterion. I will call this 'quale parsing'.

Behind HPEP there is a deeper or harder problem: it is not unique to consciousness. The 'easy' problems of HPEP such as language, perception, attention, learning and memory, are seen as explainable through direct mappings from macroscopic functional properties to microscopic neuronal ones following a strict bottom-up causality. The hard problem is declared to be special because it is assumed that the quale of experience, or Phenomenal (P) consciousness as opposed to Access (A) consciousness [9], is not amendable to a third person verifiable treatment because it is perspectival. There are at least four problems with this HPEP dichotomy beyond the obvious one that a definition of 'being like' is not proffered. First, the hard problem is not unique to conscious experience, or quale, alone but holds for all memory-dependent processes of intentional systems or mental states. Franz Brentano pointed out that mental states are intentional and directed towards something. Indeed, all goal-oriented behaviour is founded on agency and its associated mental states, experienced or not. Spinoza called this connation, in the sense that nature endeavours to persist in its own being, or conates, towards future states [10]. The uniqueness of experience results from this continuous goal-oriented and memory-dependent cycle of agent–environment interaction in the brain, body, environment nexus [11]. The upshot of this is that as soon as an internal state of an agent becomes dependent on memory, it will be perspectival, defining the aboutness of the individual agent and resisting a third person description whether it is ultimately experienced or not. Hence, the study of mind, and by extension consciousness, is pervaded with hard problems while the easy ones seem to be the exception, if there are any. Second, easy problems are hard in their own way because biological systems do not follow bottom-up causality. We can take the well-known example of the 302-neuron brain of *Caenorhabditis elegans*, fully described yet still not understood, or the role of overt behaviour itself in structuring perceptual systems via behavioural feedback [12]. Third, the HPEP dichotomy follows from a reductionist methodology, which, as argued earlier, is intrinsically dualistic. Despite its triumphs in explaining the physical world it has still not been very effective in explaining living systems (e.g. [13]). Let us consider the developing field of epigenetics that shows how macroscopic properties of phenotypes and environments in turn affect their microscopic organization [14]. We can consider the control of phenotypic variability through regulatory genes that in turn are influenced by environmental and phenotypic factors [15], the relation between maternal stress and traits of her offspring [16] or the aforementioned phenomenon of behavioural feedback that has shown that macroscopic behavioural structure imposed by acquired habits can directly lead to changes in the microscopic synaptic organization of perceptual systems [12]. In all cases, we see a complex multi-scale organization where the multiple levels of organization of brain, body and environment are tightly and bidirectionally coupled. Fourth, the HPEP dichotomy automatically leads to the implication that consciousness is a fundamental property of matter [17]. A similar consequence is echoed by recent attempts to find a quantitative consciousness index [18]. But delegating conscious experience to matter via panpsychism does not solve the problem but rather further obscures it. We have to pay heed to our history: behaviourism also sought solace in a (mis)interpretation of physics, stifling progress in the study of mind for many decades from which we still experience the aftershocks [19]. Moreover, matter itself has become a complex construct ever since Newton's introduction of forces acting at a distance, the dissociation of description from ontology starting with the Fourier transform and Maxwell's mathematical theory of electromagnetism. This epistemological divergence was further radicalized with the advent of quantum mechanics leading to what have been called 'Ghost Fields' that eliminated the plausibility of macroscopic intuitions about matter completely [20]. The putatively profound consequence of panpsychism, however, hides an explanatory nihilism [21] that answers the 'why' and 'how' of consciousness, with 'it just is', which is not very satisfactory or as William James put it: 'I can take no comfort in such devices for making a luxury of intellectual defeat. They are but spiritual

chloroform. Better live on the ragged edge, better gnaw the file forever!' [22, pp 179–180].

An alternative, and more straight forward hypothesis—which has not been exhausted yet—is that the HPEP and its terminus of panpsychism is based on a false dichotomy and that consciousness is a unique feature of living systems, the product of biology rather than a property of matter [2,23] where we must consider its function and features in terms of survival and fitness [24]. The above analysis of HPEP also suggests that we do have to change the methods of our approach from a strict reductionist model to one that can encompass the multi-scale organization of psychobiological systems. I argue that biologically grounded models of the brain, body, environment nexus provide the method to accomplish this: synthetic consciousness [11,12].

2. Distributed adaptive control: a theory of the brain, body, environment nexus

DAC is defined against the background of the fundamental challenges of both traditional artificial intelligence and connectionism [25] and the tug of war between empiricism and rationalism in the twentieth century study of mind and brain [19,26]. DAC emphasizes the epistemic autonomy of its models and realizes this following a synthetic methodology of convergent validation, which integrates constraints taken from anatomy, physiology and behaviour in one model that is embodied in a robot and situated in the real world [25]. DAC follows Claude Bernard and Ivan Pavlov in conceptualizing the brain as a control system that maintains a metastable balance between the internal world of the body and brain and the external world through action. This pertains to both the physical and the epistemic needs of the agent, e.g. the location of a food source and the exploration of an environment for map building. The question thus becomes: 'what does it take to act?' DAC proposes that an animal needs to optimize a number of objectives in order to determine a behavioural policy or procedure to achieve a goal state, or the 'How' of action. These are the 'Why' of the *motivation* for action in terms of needs, drives and goals; 'What' defined by the *objects* in the world and their affordances pertaining to needs; 'Where' of the *location* of objects in the world, the spatial configuration of the task domain and the location and confirmation of the self, as in the position of body parts; and 'When' defined by the sequencing and *timing* of action relative to the dynamics of the world and the self or the H4W problem [11].

The *H4W problem* of survival in the physical world harbours a complex set of computational challenges that brains must resolve. DAC proposes that goal-oriented action in the physical world emerges from the interplay of the processes subserving H4W [11,27]. These processes are organized in a four-layered control architecture with tight coupling within and between these layers distinguishing: the Soma (SL), Reactive (RL), Adaptive (AL) and Contextual (CL) layers (figure 1). Across these layers a columnar organization exists that at every level of the hierarchy deals with the processing of states of the world grounded in exteroception, the self, derived from interoception and action sensed through proprioception. The latter mediates between the first two via the environment. The RL is a model of the evolutionary ancient core behaviour systems (CBS) of the brainstem and midbrain [28,31,32]. The AL provides for the grounding of the representations of the world

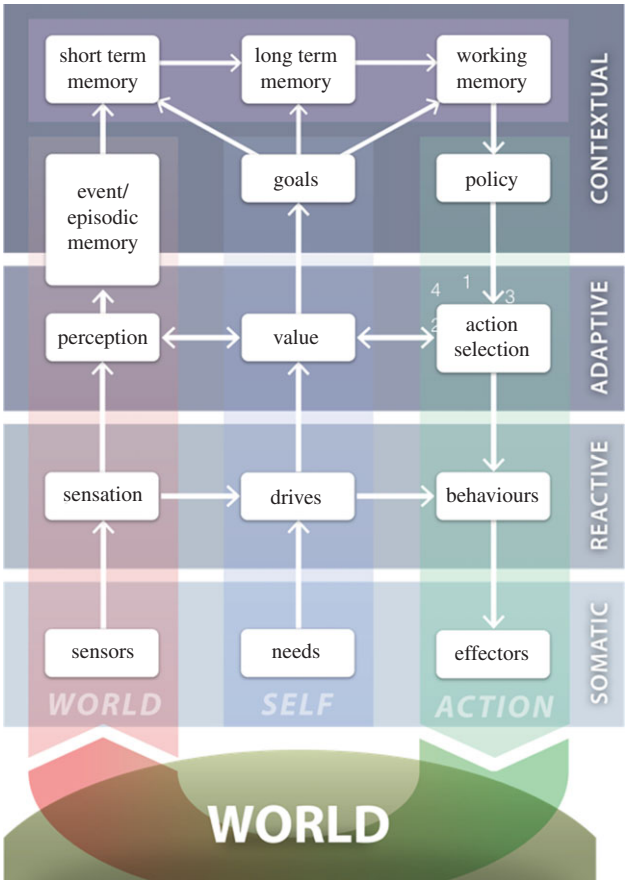

Figure 1. A highly abstracted representation of the distributed adaptive control (DAC) theory of mind and brain showing its main processes (boxes) and dominant information flows (arrows). DAC is organized along four layers (soma, reactive, adaptive and contextual) and three columns (world, self, action). The 'soma' designates the body with its sensors, organs and actuators. It defines the needs, or self-essential functions (SEF) the organism must satisfy in order to survive. The reactive layer (RL) comprises dedicated behaviour systems (BS) each implementing predefined sensorimotor mappings serving the SEFs. In order to allow for action selection, task switching and conflict resolution, all BSs are regulated via a, so-called, allostatic controller that sets the internal homeostatic dynamics of BSs relative to overall demands and opportunities [28]. The AL acquires a state space of the agent–environment interaction combining perceptual and behavioural learning constrained by value functions defined by the allostatic control of the RL, minimizing perceptual and behavioural prediction error [29,30]. The contextual layer (CL) further expands the time horizon in which the agent can operate through the use of episodic and sequential short- and long-term memory systems (STM and LTM, respectively). STM acquires conjunctive sensorimotor representations assisted by episodic memory as the agent acts in the world. STM sequences are retained as goal-oriented sequences in LTM when positive value is encountered, as defined by the RL and/or AL. The contribution of stored LTM policies to decision-making depends on four factors: goals, perceptual evidence, memory chaining and valence while action selection is further biased by the expected cost of the actions that pertain to reaching a goal state. The content of working memory (WM) is defined by the memory dynamics that represent this four-factor decision-making model. See text for further explanation.

and the self through perceptual and behavioural learning systems. Tuning of representations in somatic time becomes of greater importance when the entropy of potential sensory states increases and cannot be predicted *a priori* as in the case of distal senses such as vision and audition and when the degrees of freedom of the body expand allowing the agent a broader range of actions and their associated sensory

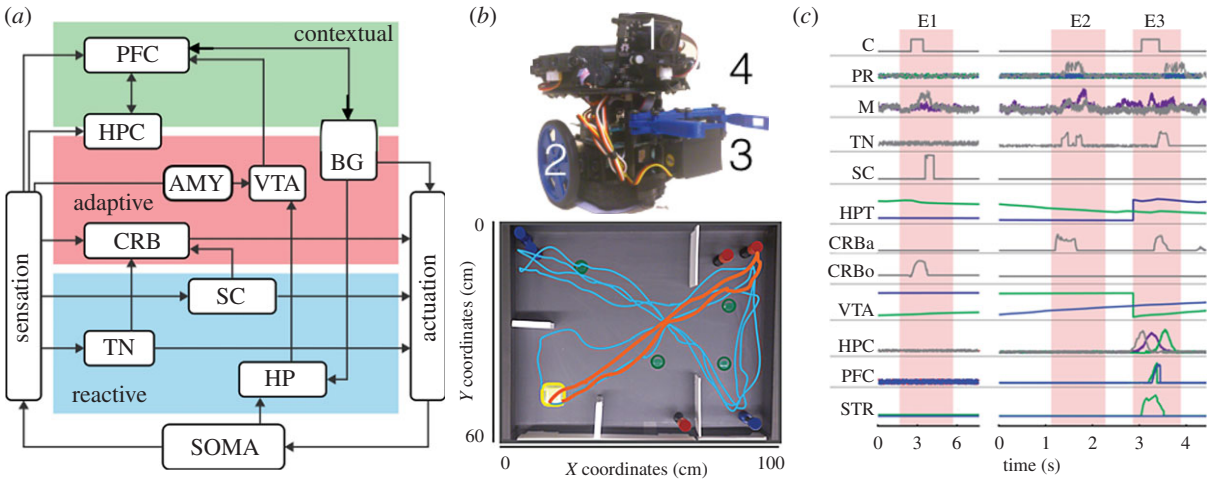

Figure 2. The DACX model maps the DAC architecture (figure 1) to main brain system and validates the resultant model in a foraging task using a mobile robot. (*a*) The reactive layer (blue) comprises midbrain/brainstem core behaviour systems (CBS) including trigeminal nucleus (TN), the superior colliculus (SC), central grey (actuation) and the hypothalamus (HP). The adaptive layer (red) includes models of the cerebellum (CRB), amygdala (AMY), ventral tegmental area (VTA) and basal ganglia (BG). The contextual layer (green) includes prefrontal cortex (PFC) and the hippocampus (HPC). All components are realised with synthetic neurons. (*b*) Top: the foraging robot (10×10 cm at its base) comprises a camera (1), wheels (2), a gripper (3) and five proximity sensors (4). The environment contains a home base (yellow), obstacles (green) and two kinds of rewards (blue and red). The robot can only consume the reward items at the home base and thus has to perform hoarding. Collisions sensed by the proximity sensors are mapped onto TN. Visual states are mapped to AMY/VTA to extract the reward quality from colour, while features extracted from the full image are processed by CRB (for the acquisition of adaptive proximal object related actions such as collision avoidance), HPC (map creation and utilization) and PFC (decision-making). The internal states of the robot are mapped onto HP. Bottom: example trajectories of the agent foraging in the environment at trial 1 (blue) and at trial 12 after reaching a stable hoarding trajectory from between the home base and a red target. (*c*) The dynamics of the neuronal control system at three distinct stages of behaviour: E1, cue-based navigation; E2, collision avoidance; E3, decision-making and reward delivery. C, colour detection; PR, proximity sensor; M, motor (left: grey, right: purple); TN, 'pain' response; SC, orienting response; HPT, internal drive state displaying two competing SEFs; CRBa, cerebellar circuits associated with avoidance responses; CRBo, cerebellar circuits for acquired approach responses. E1: Adaptive cue-based navigation, during exploration of a visual cue (C) triggers an adaptive approach response in the cerebellum (CRBa) conveyed to the motors (M) which partially overwrites the reactive orienting response triggered by SC, allowing the agent to turn towards a resource location. E2: Acquired obstacle avoidance: a proximity sensor signal (PR) allows the robot to efficiently prevent collisions with obstacles by being associated with an avoidance response (CRBo). The unconditioned stimulus, i.e. collision, associated response in the TN is partially removed due to the peripheral disruption of the US. E3: Decision-making and reward delivery, when the agent reaches the home location, visually identified by a landmark (yellow patch) as well as represented by the acquired internal spatial representation (HPC), a reward value associated with the hoarded item is delivered (STR, green) and the allostatic value for the related internal state encoded in HPT (blue), increasing its values and, consequently, decreasing the motivation to pursue that type of resource encoded in VTA (green). Activity in the VTA module directly affects the decision-making process (PFC) performed at the beginning of the trial, biasing the decision towards the most urgent need (PFC, blue).

consequences. The acquisition of the sensorimotor state space at the level of the AL is based on predictive mechanisms [30] in order to optimize encoding capacity and robustness and to counteract input sampling biases due to behaviour itself or behavioural feedback [12]. The AL models the brain systems underlying classical or Pavlovian conditioning [33,34]. The CL forms goal-oriented policies by constructing sequential representations of conjunctive sensorimotor states following a four-factor decision-making model integrating and selecting goals, perceptual evidence, memory biases and valence. The CL models operant conditioning and the DAC solution to H4 W in robot foraging tasks is both Bayesian optimal and epistemically autonomous, i.e. all representations informing decision-making are acquired by the agent in somatic time as opposed to being predefined [12].

DAC has been advanced along two classes of models to realize convergent validation. On one hand, a whole brain architecture approach was followed which facilitates the mapping to behaviour. On the other hand, components of the architecture and their basic operating principles have been linked to the mammalian brain through anatomically and physiologically constrained models [27]. These two lines of the investigation have been recently integrated into a first embodied whole brain model (DACX) comprising detailed models of core brain structures including: cerebellum, entorhinal cortex, hippocampus and prefrontal/premotor cortex (figure 2, [35]). We will later again use the DACX model to demonstrate the quale parsing methodology.

3. The dimensions of consciousness

The science of consciousness is an active field of research with a plurality of hypotheses, data, ideas and open questions that require a solid theoretical framework. A number of complementary core principles have emerged which can be summarized in the 'grounded enactive predictive experience model of consciousness' (GePe). The GePe framework identifies six principles that theories of consciousness must account for:

(i) Consciousness is grounded in the experiencing physically instantiated and environmentally and socially situated self. Experience requires a self that does the experiencing [2,8,36]. For instance, Edelman proposes primary and secondary forms of consciousness that relate to the expanding temporal horizon in which the self operates, from the instantaneous physical experience (primary) to the imagined future and remembered past (secondary). Metzinger has further elaborated this notion where the self progresses from a globalized identification with the body or first person perspective (1PP), to

a transparent spatio-temporal self-localization in the world or minimally phenomenal self (MPS), based on a form of representation of the self, to a full-fledged phenomenal first person perspective or strong 1PP (s1PP). 1PP begins as a point of convergence of sensory experience then coalesces into a strong form where the self is internally represented as reflecting the organization of the body and its sensorimotor coupling to the world (see GePe principle ii, below) followed by the representation of the object- and action-directedness of the intentional self found in the s1PP. As interactive and social dynamics are rich sources of sensory experience and feelings, they become part and parcel of the representation of the self, which is not only physically but also socially instantiated [37]. In a sense, this view on self and consciousness also reflects a trend in cognitive science to ground knowledge and experience in embodiment, situatedness and interaction [38,39]. Damasio has recently advanced a similar proposal suggesting that consciousness requires representations of self to enter into memory, essentially creating an s1PP [40]. A further variation on the role of embodiment in the nature of consciousness is the passive frame theory of Morsella that ascribes to consciousness the central role of facilitating the control of the skeletal–muscle system [41]. Challenges to this perspective are cases where embodiment is disrupted as in locked-in syndrome, phantom limb pain or for subjects born without limbs who yet experience them [42].

(ii) Conscious experience is defined in the sensorimotor contingencies of the agent environment interaction. In neuroscience, cognitive sciences and robotics a shift is taking place from a top-down representation-centred framework towards a paradigm that focuses on the intimate relation between embodied perception, cognition and action. Although many proponents have supported such an 'action-oriented' paradigm over the years, starting with Bernard, Pavlov and his mentor Sechenov [43], it has only recently begun to gain traction. In this view cognition is not a database serving planning, isolated from action and perception in a strict sense-think-act cycle but rather cognitive processes are closely intertwined with action and can actually best be understood as 'enactive', as a form of practice themselves [38,44]. The intrinsic action-relatedness of perception, cognition and experience is the core consideration of the, so-called, 'sensorimotor contingency theory' (SCT) put forward by O'Regan & Noe [45]. According to SCT, the agent's sensorimotor contingencies are law-like relations between movements and sensory inputs providing the foundations for qualia. A challenge for this framework, as for its behaviourist ancestors, is how it can scale-up to non-sensorimotor experiences without making additional assumptions on the internal processes that contribute to them. We can think of dreams or when the interfaces to the body and its sensorimotor contingencies are disrupted as in the case of locked-in syndrome.

(iii) Conscious experience is predicated on predictions. The idea that perception and cognition are defined by predictive models of environmental causes of sensory input and the outcomes of action enjoys a rich pedigree extending back at least as far as Plato who already struggled with the issue of beliefs as predictions in his *Theaetetus* (369 BC). In the modern era, it goes back to von Helmholtz [46] and the seminal work of Tolman [47] and Craik [48]. Indeed, sensorimotor contingencies do not only exist instantaneously (GePe principle ii) but can also be predicted by virtue of their invariant relations in time [49]. Indeed, temporal invariance underpins the Humean definition of causality also expressed in the principles of classical conditioning where the conditioned stimulus is inferred to 'cause' the unconditioned stimulus. It has been proposed that cognition and consciousness can be based on internal simulations of the possible scenarios of interaction with the world using, so-called, forward models (e.g. [50,51]). It is through simulation that an 'internal' world can appear in consciousness, freeing the organism from its immediate physical environment, creating a self-generated virtual reality [52]. Merker has argued that this simulated internal world compensates for the uncertainties in the interpretation of sensory states due to self-induced motion [53].

That the brain is organized around prediction has been reaching recent prominence in the so-called 'Bayesian brain' and 'predictive coding' frameworks anticipated by Dom Massaro in his analysis of multimodal speech perception [54]. In this view, core structures of the brain are engaged in hierarchical Bayesian inference, extracting generative models of both sensory inputs and the consequences of action by reducing surprise or 'free energy' [55], where in neurophysiological terms 'top-down' connections convey predictions, while 'bottom-up' signals convey prediction errors [29,49]. Indeed, the predictive coding perspective currently basks in a growing popularity (for a summary, [56]). However, this popularity does not shield the predictive coding approach from a number of challenges. For instance, from spinal cord to frontal lobes, the brain is structured around a variable range of prediction-based systems many of which operate outside of the reach of consciousness, most notably the cerebellum, which comprises about 70% of neuronal volume [57]. It is thus not clear which specific property of predictive processing pertains to consciousness. For instance, some correlate of this view has been reported in coma patients in terms of the absence of top-down prediction modulation [58], while patients with severe deficits of consciousness can elicit anticipatory actions to the occurrence of a stimulus in classical conditioning [59]. Furthermore, it is unclear what the relations are between probabilistic representations required by a Bayesian brain and the singular and unitary character of conscious experience itself. Merker argues that while in cortex information is generally maintained in the form of probability distributions, the content of consciousness is instead linked to the 'collapsing' of probability functions into a simpler unitary format required for subcortical processing [53], providing global best estimates of variables of interest within narrow time windows [60]. In addition, the question is whether conscious experience itself is dominated by top-down predictions or bottom-up prediction errors, which we will revisit below. A last challenge to the predictive coding framework is that it is normative and it does not inform us directly on how brains actually can realize the optimizations

that they 'should' perform.

A further variation on prediction-based theories of consciousness is the attention schema theory [61]. In this case, underlying consciousness is the process of attention, which an observer initially attributes to an outside agent in order to infer subsets of sensory information of relevance to that agent, allowing the observer to adapt their own actions. Consciousness is then seen as the results of the ascription of such attentional states to the self.

(iv) Conscious experience optimizes both information differentiation and integration. Following Edelman it is a deeply significant fact that each and every conscious scene is both highly *integrated* and massively *differentiated* [2]. Tononi has further formalized this idea through a range of increasingly more complex complexity measures culminating in Integrated Information Theory (IIT) and its 'consciousness index' Φ, expressed in bits [18,62]. As a dynamical measure of network complexity, Φ seeks to operationalize the intuition that complexity of brain functions results from simultaneous differentiation and integration of information by the underlying neuronal circuits and signalling processes. Differentiation refers to the existence of specialized neuronal populations with distinct functionality, while integration results from the coordination among these populations leading to the emergence of coherent cognitive and behavioural states and consciousness. This combined process creates patterns of high complexity due to its causal dynamical interactions, over and above the information generated independently by the disjoint sum of its parts. However, to map this intuition into a coherent practical measure is less easy than it looks and especially it faces challenges with respect to scaling and the nonlinear properties of brain networks [63,64]. Conceptually, IIT shares properties with, so-called, higher order theories (HOT) of consciousness that propose that experience depends on the access to low-level brain states via high-level meta-cognitive ones [65].

The question facing IIT, however, is whether information integration and differentiation as defined by IIT is specific for consciousness. IIT redefines consciousness as having a non-zero value on Φ. If we assume that subconscious processes are probabilistic while qualia are unitary, one can predict that subconscious states must have a higher value of Φ. Indeed, many systems that would appear non-conscious satisfy this criterion, e.g. a $N \times N$ grid of XOR gates has a Φ of $\sqrt{N}$ (see [66], for a critical analysis and summary). In this respect, IIT is facing a variation of the frame problem by basing its analyses on an exclusion principle: how many informational states does the system reject in an unbounded informational space. Indeed, by elevating a method into a theory, IIT makes counterintuitive ontological commitments: in the IIT universe, a photodiode has 1 bit of consciousness [67]. Panpsychism is the logical consequence the proponents of this methodology have gravitated towards. For research purposes and clinical applications measures of the dynamical complexity of the brain are of great assistance. However, this does not imply that such measures constitute a theory of how the mind/brain actually works or provides a hypothesis on the function of consciousness. If the consequence of Φ is panpsychism, where consciousness becomes a property of matter generated at the big bang rather than a feature of biological systems produced during evolution, the cost of it might be too high relative to its intellectual benefit.

(v) Consciousness depends on both highly parallel, distributed implicit factors and metastable, continuous and unified explicit factors. Theories of consciousness are tightly constrained and informed by evidence regarding unconscious processing [68]. In Baars' 'global workspace' (GW) theory specialized unconscious processors compete for access to a central resource: the conscious global workspace. In this view consciousness is ascribed to content that is received from and broadcast back to a broad network of unconscious modules or processors. In this way, consciousness provides a serial and integrated stream of qualia that is produced by many subconscious processes. The key parameter that defines whether content becomes conscious is the ability to penetrate many of these processors. In this respect, GW is an example of, so-called, access-consciousness [69]. The integration and serialization provided by GW provides for behavioural flexibility by allowing unconscious processors to generate fast responses in familiar situations, while in novel situations, the integrated quale broadcast from GW can facilitate the production of new responses [68]. The, so-called, global neuronal workspace (GNW) hypothesis proposes that the workspace comprises perceptual, motor, attention, memory and value areas that form a common higher level unified information space that serves a similar role as the GW, largely defined through the specific anatomy of cortico-cortical projections [3]. GW/GNW is essentially a proposal on how content is generated for conscious experience; the main function ascribed to the GNW is that of assisting in problem solving and executive control [70]. GW/GNW is a modern neurocognitive version of the Freudian notion of subconscious factors driving experience and action. The challenges faced by GW/GNW are that it is a proposal on how qualia can be generated; it is unclear, however, how the experience of this content is realized. For instance, how is selection across the processors controlled, are conflicts resolved and why would conscious experience have to re-penetrate the subconscious processors? In addition, it is based on the assumption of resource limitation driving the realization of serial conscious experience but does not declare what that resource is.

(vi) Consciousness is decoupled from real-time performance. Whereas at least since Homer we have placed experienced agency at the heart of our existence, which was elevated by Descartes to the modern dogma of dualism, a number of converging lines of evidence show that humans are largely unaware of the causes of their own thoughts and actions (see [71] for a review). The interpretation that consciousness is an epiphenomenon seemed a forlorn conclusion, i.e. an evolutionary leftover with no operational relevance [72]. Indeed, many cognitive processes can be performed without reportable awareness of the relevant stimuli or contingencies, and some processes (e.g. overlearned motor responses) are reported to be more effective when realized by subconscious systems ([68,73], see [74] for a criticism). Less appreciated but equally fundamental, is the notion that motor actions and intentions can both be

subconscious as well as conscious [75,76], that subconscious intentions reliably precede conscious awareness of motor actions [77–79] and that behavioural goals can be set by subconscious factors [80]. Another category of implicit factors in experience and action are emotions. Indeed, emotion and consciousness are tightly coupled and conscious experiences generally involve affective components, both transiently (e.g. delight, surprise) and as a background mood (e.g. sadness, contentment, anxiety) [81]. Since James-Lange it has been suggested that emotions arise as perceptions of bodily states [82] and that autonomic signals can reflect implicit reactions to salient stimuli, including prediction errors [83]. Further, it has been argued that the processes underlying volitional behaviour, such as implicit learning, evaluative conditioning and subconscious thought are intrinsically goal-dependent requiring forms of attention, while operating outside of awareness [84]. In all cases, conscious and subconscious processes are closely coupled and interact strongly in generating the stream of consciousness and adaptive behaviour [85] as also expressed in GW theory. They can be seen as complementary since unconscious processing can be sensitive to patterns, regularities and other structures within signals substantially prior to conscious awareness, suggesting that the content of consciousness is biased and based on subconscious factors [68,86].

A number of dual-process theories have been advanced, which aim at disentangling the relationship between subconscious and conscious processes. Example models are the notion of fast automatic and slow deliberative processes in decision-making [87] and the distinction between reasoning, planning and monitoring, where the latter process only has indirect access to mental states and is forced to, in turn, interpret and speculate on the causes of action the agent generates [88]. Dual-process theories face the fundamental question of how these processes are maintained in isolation and how the exchange of information between them is regulated. In particular, the question looms whether these multiple processes are in turn coherent or descriptions of a further heterogeneous set of subsystems possibly leading to an infinite regress [89]. Hence, any theory on consciousness must take a stance in the free will versus epiphenomenon dilemma. In particular, the defence of a causal role of volition in action will have to grapple with the delay at which conscious experience occurs relative to the real-time generation of action as observed by Libet [78]. This remains a fundamental issue even though the exact value of the delay between the subconscious physiological process of action initiation, or the readiness potential, relative to overt action has been under some scrutiny [90].

4. Defining consciousness

The renaissance of the science of consciousness has not generated an accepted definition. One current trend is to bypass the definitional stage and settle for an operational one as in IIT, while a second approach is to follow the NCC route. Definitions, however, are helpful in shaping science. Or rather, if we define science as advancing third person verifiable descriptions, we must be willing to state what we are describing beyond intuitive suggestions such as 'what it is like' to be awake or a bat. I will base the definition of consciousness on an analysis of deficits of consciousness, in particular, hemispatial neglect grounded in considerations on the neuronal substrate of consciousness.

(a) The neuronal substrate of consciousness

The neuronal correlate of consciousness (NCC) has become the focus of research for the last three decades [1]. It is common to distinguish between the level of consciousness versus its content or awareness ([91] but see [92] for a critical analysis). The level of consciousness is defined from wakefulness with high vigilance and arousal to coma and general anaesthesia, while awareness defines the content of consciousness. The neuronal architecture underlying these two dimensions of consciousness can also be described as comprising three main levels: core behaviour systems (CBS), brain activating system (BAS) and the fronto-parietal thalamo-cortical system.

Bjorn Merker who advocates a prediction for stabilization of motility hypothesis (GePe iii), has suggested on the basis of a number of behavioural, anatomical and clinical observations including the analysis of hydraencephalic children, that the integration upon which conscious states depend can be traced back to the midbrain zona incerta, a hub in the interaction between brainstem and cerebral cortex [31,93]. In further elaborating the centrencephalic theory of consciousness of Penfield & Jasper [94], he proposes that primary consciousness resides in the CBS of the midbrain comprising the superior colliculus at the roof representing sensory states, the hypothalamus at the floor conveying states of the agent and the intermediate structure, zona incerta, being a conduit for their interaction and action generation via the periaqueductal grey [31]. Indeed, the zona incerta has also been implicated in deficits of consciousness such as epilepsy. Others have made comparable proposals on the neuronal substrate of primary or proto-consciousness [95,96] that all follow the world-self-action triade of the RL of DAC. A detailed analysis of both the vertebrate and invertebrate nervous system has led to the suggestion that the central complex in the insect brain serves the same functions as the CBS and thus provides an anatomical signature for invertebrate consciousness ([97], but see [98] for an alternative interpretation of this structure). An associated proposal advances the hypothesis that primary consciousness and the CBS evolved in two stages about 520 Ma ago, driven by the emergence of distal sensing afforded by vision and the required need for the realization of map-like representations [99]. This, in turn, further accelerated vertebrate evolution by allowing the exploitation of more complex niches. Further support for the role of phylogenetically ancient structures in the genesis of consciousness are studies based on the impact of anaesthesia on the human brain, which point to a key role of BAS (e.g. locus coeruleus), hypothalamus, thalamus, anterior cingulate (medial prefrontal area) and connectivity in frontal-parietal networks, thus linking CBS and BAS [100]. Thus, multiple lines of evidence point to the brainstem/midbrain CBS as providing a substrate for the primary consciousness of instantaneous feeling or sentience. It is important to note that the CBS satisfies all components of the GePe framework: embodiment, sensorimotor contingencies, prediction and integration [31], while we can observe that the distributed competitive dynamics of the superior colliculus combined with its multi-modal integration will realize a global workspace compatible mode of operation [101]. This midbrain primary consciousness system adds an additional significant

feature by interfacing to the neuromodulatory activating systems of the evolutionary ancient midline arousal systems that modulate the state of forebrain structures.

Conscious is commonly defined in terms of its absence as in coma and vegetative state. These deficits of the level of consciousness are mostly related to lesions to the midbrain and BAS discovered in the 1940s, or the ascending arousal system. A recent study that monitored the activity of the brain while it is returning to wakefulness showed that first the neuromodulatory centres of the brainstem became active followed by the hypothalamus, thalamus and the anterior cingulate cortex (ACC) [100,102]. Indeed, lesions to BAS induce sleep-like states in the thalamus and neocortex [103]. Nuclei comprising the BAS project globally to the thalamus (via its dorsal branch) and the neocortex (though the ventral branch which includes cholinergic basal forebrain projections) controlling the global state of arousal using a wide range of neurotransmitters. BAS includes the monoaminergic noradrenergic locus coeruleus, the dopaminergic neurons of the ventral tegmental area (VTA) and the substantia nigra, the serotonergic raphe nucleus (which also contain some dopaminergic neurons), the pedunculopontine and laterodorsal tegmental cholinergic nuclei and the histaminergic tuberomammillary nucleus [104]. These systems all contribute in various ways to sleep, wakefulness, arousal, behavioural activation and value processing. BAS can thus be seen as the key regulator of the level of consciousness and the question is to what extent it also modulates its content and/or conveys this content to subconscious systems?

Minimally conscious state (MSC) patients still display activity in their thalamo-cortical system suggesting that residual awareness persists. Indeed, patients who appear vegetative can display physiological states of their neocortex similar to those shown by healthy subjects [105]. A dramatic demonstration of the role of the thalamo-cortical system in consciousness was that bilateral stimulation of the central thalamus could partially restore behavioural reactivity in a minimally conscious patient [106]. The level of consciousness depends on dynamic changes in the thalamo-cortical system also due to intrinsic firing properties of thalamic neurons which can switch between tonic and phasic firing modes [107]. In these thalamic neurons, hyperpolarization-induced low frequency bursting can play a key role in pathological thalamo-cortical dysrhythmia (TCD) observed in Parkinson's disease, neuropathic pain and tinnitus [108]. This hyperpolarization is, in turn, the result of a lack of inhibition onto the subthalamic nucleus due to pathological changes to the basal ganglia. It has been suggested that the thalamus itself comprises neurons that project to the neocortex realizing a function in their target regions comparable with that of the BAS [109]. The firing properties of these neurons of central thalamic nuclei are, in turn, controlled by a frontal cortical–striatopallidal–thalamo-cortical loop. Interestingly, enough this is fully consistent with the general circuit principle of TCD [110]: lack of medium spiny cell inhibition to the internal segment of the globus pallidus leads to hyperpolarization and pathological low-frequency bursting in thalamic cells, switching off the corresponding cortical circuit. The central role of the thalamus in the regulation of consciousness is further supported by the widespread thalamic neuronal cell death observed in vegetative state patients [111]. Another example, which illustrates the central role of the thalamo-cortical system as a core component of the neuronal substrate of consciousness, is that vasoconstriction of the middle cerebral artery, which provides blood to both cerebral cortex and the thalamus, can result in a transient loss of consciousness (syncope), associated with an increase of slow wave activity in the EEG indicative of functional inactivation of the cortex [112]. Also, TCD is accompanied by a characteristic downward shift of the power spectrum [108]. A similar pattern can occur in epilepsy where pathological dynamics can be observed in fronto-parietal networks and the basal forebrain, thalamus, hypothalamus and upper brainstem, also dubbed the consciousness system [113]. Interestingly, an analysis of the distribution of activity patterns during seizures where consciousness was absent, showed a distinct increase of activity of the fronto-parietal system in the delta range (1–2 Hz) correlating with seizure activity in the temporal lobe [114]. This low frequency is usually associated with the absence of consciousness in sleep-like states, coma and anaesthesia. A similar distinct dynamic pattern, binding fronto-parietal networks distinguished patients in a coma from those suffering from MSC, where the former showed enhanced delta power and reduced alpha and theta when compared with the latter. Hence, the thalamo-cortical fronto-parietal system appears to be of central importance in defining both the level and content of secondary consciousness. The link between thalamic dynamics and TCD shows that in physiological terms the level of consciousness can be regulated in an anisotropic fashion, where sub-circuits of the thalamo-cortical system can be switched on or off dependent on the dynamics of thalamic projections.

Deficits of consciousness are not necessarily expressed as changes to its overall level. Stroke patients can suffer from specific deficits of consciousness as in hemispatial neglect where they are not considering part of the extra-personal (visual/auditory), personal (haptic) or representational (imagery) task space contra lateral to the lesion in their overt actions and/or verbal reporting [115]. It is the result of a number of non-exclusive deficits that can appear in combination including: directional deficits in the control of attention, representation of space and/or action directedness and non-spatially lateralized mechanisms as in problems with sustained and selective attention, a bias towards local features in a visual scene, as well as a deficit in spatial working memory [116]. Neglect is mostly associated with a lesion-induced reorganization of the neocortical frontal-parietal networks of the right hemisphere (RH) [117,118]. It can be seen as a network deficit due to diaschisis [119] showing a slowing of the EEG [120] that can be explained in terms of TCD [121]. Interestingly, about 82% of RH patients and 62% of left hemisphere (LH) stroke patients suffer from spatial neglect during the first days after stroke. About 20% of these patients, mostly LH, recover spontaneously between three to eight months post-lesion, while about one-third of RH patients struggle long-term with spatial neglect. Neglect patients commonly show an anosognosia, insisting that their experience of the task space is complete and unitary including at the level of their autobiographical memory [122,123], demonstrating that conscious experience is integrated and coherent even in the absence of perceptual evidence.

Spatial neglect offers a unique possibility to investigate consciousness as it pertains to its content (figure 3). As an example, we can consider a visual detection task, where right hemispheric stroke patients (RH) with neglect were asked to detect local or global targets that appeared in random positions (figure 3*b*). The overall detection rate significantly declines with an increasing set size when compared with controls (figure 3*c*, left column; $p < 0.05$), while the reaction time also increases about threefold (figure 3*c*, right column; $p < 0.05$). This reduction in performance is most pronounced when the target is defined as an

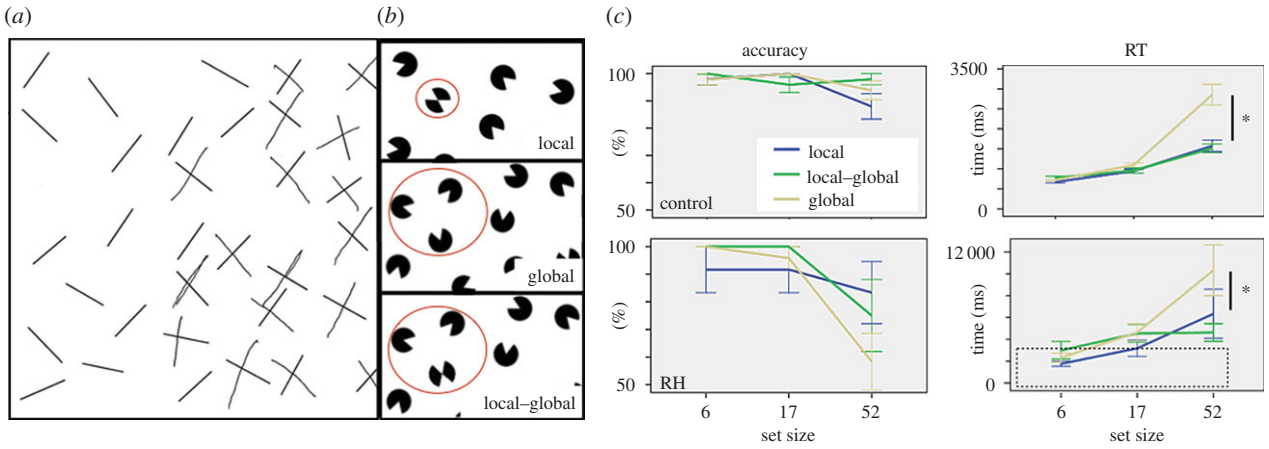

Figure 3. Hemispatial neglect. (*a*) Standard drawing test performed by a patient (male, 64 years) with a right haemorrhagic stroke showing a reduced ability to fill the left side of a workspace with crosses. (*b*) A visual search task where subjects have to detect a local (top), global (middle) or combined (bottom) distractor in a visual display, which occurred with 50% probability in 72 trials. The distractors are indicated with a circle. The global distractor is a Kanizsa triangle. (*c*) Performance of control subjects and right hemispheric (RH) stroke patients in terms of accuracy (left column) and reaction time (RT; right column). For visibility, the RT of the control group is displayed in a reduced range, which is reflected with a dashed box in RT plots for RH. Asterisks indicates a significant difference at $p < 0.01$. $N_{control}$: 10; N_{RH}: 8. Adapted from Campillo *et al.* [124]. See text for further explanation.

illusory contour, where we observe a doubling of the reaction time as compared to the other stimulus conditions. This suggests a disruption of top-down processing, as this condition requires the binding of multiple local features. Interestingly, in this experiment, group RH underperformed the healthy controls in both target present as well as target absent trials, e.g. reporting both false negatives as well as false positives (data not shown). This surprising reduction in performance suggests changes in search strategy and or diminished perceptual confidence [125].

The neglect patient example illustrates that the stream of consciousness is *unitary* and *continuous* ([22,126], GePe iv). It can be considered *integrated* (GePe iv) *autonomous* and constructed or *virtual* at both the instantaneous and autobiographical level as also evidenced by anosognosia in neglect (GePe iii). In addition, conscious experience is *transient* in the sense that it can be reversibly disrupted through lesions to the thalamo-cortical system and BAS and anaesthesia [127]. As a result it has *levels*, controlled in a two-stage fashion through BAS and the thalamus where the latter's influence is highly anisotropic as evidenced by TCD allowing the fashioning of a GW (GePe v). A last distinguishing feature of conscious experience is that it occurs with a significant *delay* relative to the real-time action of the agent (GePe vi). It is this last feature of consciousness that defines it as a memory process and precludes it from being the cause of instantaneous action and thought.

The level of consciousness depends especially on the phylogenetically ancient structures of the BAS while it appears that its content is defined in two stages: primary consciousness or sentience through the CBS of the midbrain and secondary consciousness by means of the thalamo-cortical system of the fronto-parietal network. As described above, this organization suggests a three-layered architecture which is phylogenetically highly conserved, appears to have its origins in the Cambrian and shows similarities between vertebrates and invertebrates. We can observe two additional and significant properties of these complimentary systems underlying the level and content of consciousness. First, BAS anatomically overlaps with the CBS of primary consciousness, suggesting that primary consciousness, in turn, can regulate the level and content of secondary consciousness maintained in the thalamo-cortical system. Given the topology of the BAS ascending projections, we can also speculate that the CBS can modulate the content of secondary consciousness by selecting specific (sub)modalities over others. Second, the majority of BAS subsystems also comprise a significant part of the brain's value systems that control the neuronal dynamics and synaptic plasticity underlying perception, cognition and action. For instance, the basal forebrain cholinergic system controls the reorganization of neocortical maps in classical conditioning [128], histamine facilitates performance on complex cognitive and memory tasks [129], the serotonergic system shows dissociable effects from the dopaminergic one on problem solving strategies [130] and the noradrenergic system is critically involved at different stages of information processing and memory, in particular, in the prefrontal cortex [131]. Interestingly, the neocortex in turn controls the activation of BAS and its valuation functions, for instance via the valence hub of the amygdala [132]. In addition, a recent meta-analysis has shown that the distribution of serotonin receptor subtypes is disrupted in the prefrontal cortex in schizophrenia [133], lending further support for the notion that systems of primary and secondary consciousness are coupled both in terms of their information exchange and their regulation via BAS.

Given the above analysis we can define consciousness as a *transient autonomous memory*, which maintains a *delayed*, *virtualized*, *unitary* representation of the agent–environment nexus. This representation can be decomposed along the three columns and layers of the DAC architecture. These dimensions of experience become increasingly more dependent on the individual history of the agent captured in the memory processes of the AL and CL, combining into a 'remembered present' [2], recollected past and anticipated future of the embodied autobiographical self. Qualia are defined through contributions of these different domains that will be actively selected on the basis of goal setting. Below, I argue that this consciousness memory system supports the *normative valuation* of performance in a world pervaded with hidden states, largely due to the presence of other agents, complementing and optimizing the parallel real-time control of action. Hence, the function of consciousness is to extract hidden norms from multi-agent and social environments and to use these to optimize the future parallel control of perception, cognition and action.

(b) The distributed adaptive control of consciousness and the H5W problem

The main perspectives on consciousness summarized in GePe and its neuronal organization might appear rather heterogeneous. However, they can be seen as each highlighting specific and complementary aspects of consciousness that must be brought together into one synthesis. My goal is to define this synthesis based on the definition obtained by analysing hemispatial neglect and the distributed adaptive control theory of mind and brain. The question that will drive the hypothesis on the function of consciousness advanced here is: what is the contribution to fitness of a delayed unitary virtualized conscious experience? Here I advance the hypothesis that consciousness is a necessary ingredient of a behavioural control architecture that is able to optimize action in a multi-agent world solving the so-called, H5W problem [27].

The DAC solution to H4W described above already comprises all elements of GePe. More specifically we can observe that DAC is:

(i) Grounded in the experiencing physically embodied self: the somatic layer (SL) constitutes the foundation of the embodied brain.

(ii) Co-defined in the sensorimotor coupling of the agent to the world: the RL and AL both establish immediate sensorimotor loops with the world, the former predefined the latter acquired. Acquired integrated sensorimotor states form the representational building blocks of DAC's cognitive and higher level representational processes. We have shown how this realization of conjunctive representations occurs in the hippocampal memory system [134], which has been further experimentally confirmed [135].

(iii) Maintaining the coherence between sensorimotor predictions of the agent and the dynamics of the interaction with the world: the AL relies on prediction-based systems for both perceptual and behavioural learning to optimize coding, enhance robustness and solve behavioural feedback [30]. The memory systems of the CL operate on a combination of forward and feedback models operating at varying timescales. Indeed, DAC is an early example of a embodied epistemically autonomous predictive coding architecture [12,29].

(iv) Combining high levels of differentiation with high levels of integration: the AL and, in particular, the CL, integrate across all sensory modalities and memory systems and provide selection mechanisms to define a unique interpretation of the state of the world and the agent, generating specific actions that are optimal in Bayesian terms based on its four-factor decision-making model [136].

(v) Exploiting a global workspace at the level of the CL integrating memory-dependent implicit biases and perceptual evidence with goal states and value supporting optimal decision-making. Task relevant states are 'ignited' by the confluence of perceptual, motivational and memory evidence to form the dominant states of the CL working memory system, driving goal-oriented action [136]. The decision-making dynamics of CL thus shows and anticipates a dynamic signature believed to be specific for the GNW [137].

The analysis so far has shown that GePe at best captures necessary conditions of conscious experience and its underlying processes and mechanisms because DAC so far lacks consciousness as defined here. Looking at the concrete DAC example allows us to assess what is specifically missing: the ability to maintain a transient and autonomous memory of the virtualized agent environment interaction that captures the *hidden states* of the external world, in particular, the H5W capability of other agents, and the non-observable *norms* that agents implicitly convey through their actions. Indeed, DAC in its answer to H4W, is solely dealing with the interaction of an agent with its physical world. However, the Cambrian explosion of about 550 Ma created environments that were dominated by one more critical factor demanding a specific objective function: other agents or 'Who'. The resulting move from the H4W to the H5W problem leads to a fundamental change in information processing: parallelization, reciprocity and hidden states. Reciprocity results from a behavioural dynamic where the agent is now acting on a world that is in turn acting upon it. The states of other agents that are predictive of their actions, however, are hidden and agents must develop a capability for mind-reading based on a theory of mind [37,138,139]. At best these intentional states of other agents can be inferred from incomplete sense data such as location, posture, vocalizations, etc. or their social salience [140]. As a result the agent faces the challenge to unequivocally assess, in a deluge of sensor data and in parallel, those observable and inferred states of other agents that are relevant to its own on-going and future actions. In addition, it must deal with the ensuing credit assignment problem to optimize them. Thus, mind-reading contributes to the credit assignment problem consciousness needs to solve, it is not the solution. Rather the solution to survival in this only partially observable intentional world entails assessing: (i) what the relevant (hidden) states of the world and its agents are, (ii) what the relevant states of self are, (iii) what the objectives and norms are that other agents follow and (iv) which specific state of the agent, e.g. action, from a large repertoire of possible states, gave rise to desirable and/or undesirable outcomes given the hidden norms of the social world. I propose that consciousness is a necessary component of control systems that solve this H5W problem. More specifically, from the perspective of DAC we can detail the specification of a control architecture that solves H5W as follows:

(i) *Autonomous virtualization memory.* The hidden states of the external world, largely due to other agents, and the internal milieu of the agent are identified through internal simulations that must be maintained over extended periods of time and also in the absence of sensory stimuli. As a result, action takes place in an augmented mixed reality where sensor data reflecting physical sources of stimulation and inferred intentional states are merged and tested against both the world and internal models. This augmentation cannot take place in relation to the physical world, as sensor states do not necessarily and directly present the relevant information such as intentions of other agents. The mixed reality thus requires a dedicated autonomous memory system that virtualizes the self and the world. It is autonomous because it must be maintained in the absence of sensory information and/or when sensory information deviates from internally generated predictions. Consciousness has been described as a brain-based virtual reality (GePe iii); the current proposal casts it in terms of an augmented mixed reality where internally generated models are merged with sensory states derived

from action in the physical world. The virtual reality case would by definition be isolated from the physical world, which would be counterproductive, since action needs to be generated in the physical world.

(ii) *Parallel multi-scale operations.* Given the number of variables to be considered in a complex multi-agent world, the real-time predicaments of living systems and the finite operating powers of physical systems, i.e. brains, there is strong evolutionary pressure on implementing components for internal simulation and virtualization through parallel operations. In addition, all real or imagined agents in the environment must be tracked in real-time, defining a further functional need for parallelization. Indeed, parallel processing is one of the characterizing features of social brains, for instance from the mushroom bodies of wasps and bees to the hippocampus of vertebrates, which have been suggested to be homologous structures [98]. In the vertebrate case we can add the cerebellum. The cerebellum is credited with controlling real-time action and comprises in humans about 15 million parallel segregated loops constituting about 70% of the neuronal volume of the brain [141]. The cerebellum can be removed without apparently affecting consciousness directly [142] although it can modulate its content as in visual illusions (e.g. [143]). Moreover, the majority of cerebellar circuits project to the frontal cortex suggesting a role in the realization of internal models [144,145].

(iii) *Serialization and unification.* The agent and its physical instantiation by necessity can only commit itself to a single action at each point in time, realized through the position and confirmation of a singular body. As argued above, these actions are based on massively parallel simulations of possible states of the (social) world and self, supporting real-time inference in an intention-laden world and body. This creates a fundamentally new credit assignment problem: which value and/or future action should be assigned to which property of the real or imagined world given the outcome of a single act? If the causes of the actions of other agents are hidden, so are their outcomes and valuation, i.e. a social world by necessity operates on norms that are hidden. Indeed, this fact has led to the hypothesis that the human mind is specialized in norm extraction [146]. Hence, in order to optimize real world action the agent is forced to appraise performance with respect to a singular interpretation of the mixed reality probabilistic states that have given rise to its actions.

(iv) *Consciousness serves intentional valuation and norm extraction.* Consciousness solves credit assignment for the parallel real-time control systems that drive a single social agent by inferring norms and transferring them as value. In DAC terms, we can rephrase this as the challenge of finding alignment between the parallel and probabilistic controllers that dominate real-time interaction with a social world and the singular, virtual and serial model of that interaction. As argued earlier and demonstrated by the DAC robot-based control models, the real-time control of action requires parallel processing further amplified by the predicament of solving H5W. Consciousness is a necessary counterpart to such a real-time parallel control system: a highly integrated virtual and autonomous sequential process that runs adjacent to and integrates across both states generated by the many parallel unconscious processes and those observed in the world. This serves the extraction of norms and the valuation of the goal-oriented performance of the agent. Norm-dependent error detection projects value to the parallel action and simulation controllers, optimizing their future performance. In this way, the cooperation between parallel real-time real-world bound unconscious and delayed serial conscious control assures operational coherence through the reinterpretation and optimization of unconscious parallel loops. In order to realize this function, the process of consciousness requires an autonomous memory system that maintains the virtualized, serialized and unified description of the interaction between the agent and its social world irrespective of the signal level information the agent is exposed to [147].

(v) *The intentionality prior.* In order to bootstrap the semantics of the simulations of the hidden states of the embodied self and other agents, they are anchored in an intentionality prior, or pervasive intentionality, where novel states are treated *a priori* as being caused by agents [147]. The problem of unifying the optimization of subconscious parallel fast control is thus resolved through shifting the representational framework from signal (world) based to intention (self) based, or an intentional stance [148]. This implies that intentionality detection is operating already at the level of the RL. As an example, we can consider stop signals conveyed to a dancing honeybee through a head butt by a fellow bee in order to prevent the dancer from indicating a location that the observing bee has found harmful in the past [149]. The further interpretation of intentional cues detected in the world, or ascribed to it, capitalizes on a self as other process, or in terms of Merleau Ponty '*ap*-presentation' [150], which implies that the self and world columns of the DAC architecture are tightly coupled. Indeed, as Dan Sperber famously stated: 'the attribution of mental states is to humans what echolocation is to bats' (quoted in [151, p 207]). The self-model continuously serves as a reference for intentional cues detected in or projected onto the world and its real or imagined agents. In this case, optimizing performance also implies the ability to suppress this intentionality prior, i.e. to learn to differentiate the physical world following laws of physics from the final causes of intentionality dictating the actions of other agents. The DAC theory of consciousness (DACtoc) thus proposes that intentionality is a third Kantian prior, beyond space and time. Whereas, space and time can be defined implicitly by virtue of being embodied, the intentionality prior will require more elaborate memory specification because it is not directly observable. DACtoc thus proposes that we discover the physical world by stripping it of its automatically ascribed intentionally.

For this hypothesis to hold, BAS must be coopted by systems involved in social cognition. It is relevant to observe that value systems of the brain and their targets are strongly dependent on error-based learning. For instance, in the case of the cerebellum, learning is regulated through an error signal generated by the inferior olive [152]. This feedback loop is subject to dopaminergic projections from the substantia nigra compacta that

terminate in the deep cerebellar nuclei [153] that in turn are linked to error and novelty detection [154] and it has been suggested that these signals can serve to directly modulate the balance between feed-forward and feedback control in the cerebellum (Verschure, Herreros and co-workers in [145]). Similarly, all systems of BAS are known to respond in a similar fashion to surprise, novelty and error [131]. More importantly, the medial prefrontal cortex can in turn modulate the value-based responses of BAS, for instance through a projection to the basolateral nucleus of the amygdala, the hub of emotion regulation [155]. Specifically, a shift of activity from amygdala to frontal areas (subgenual ACC and the right temporal pole) has been reported corresponding to decisions that were either fear driven or goal oriented and fear inducing [156]. Another case in point is the VTA and its widespread dopaminergic projects to, among others, the neocortex and hippocampus, modulating memory in both systems [157,158]. VTA is itself part of a recurrent loop with the ACC, which in particular targets the VTA neurons projecting to the nucleus accumbens [159]. In addition, inputs are received from the amygdala and components of the BAS: hypothalamus and raphe nucleus. That these reward pathways can actually provide the substrate for social decision-making, norm extraction and processing is suggested by a number of studies (see [160] for a review). For instance, when human teachers are instructing learners, activity in the ACC of the teacher was shown to covary with the prediction errors of the learner, i.e. tracking the learner's performance [161]. Similarly, violations of a promise in a game theoretical set-up correlate with significant activity in ACC [162], while the same structure enhances its activity in proportion to reported envy [163]. In addition, this social value and norm processing system also overlaps with the fronto-parietal network [164], which also is home to the mirror neuron system that supports social perception and action [165]. Hence, it appears that the reward pathways of the brain are also engaged in social decision-making, multi-agent modelling, valuation and norm extraction by being coopted by frontal cortical areas and their associated subcortical systems. States of this broader norm extraction network can also enter the fronto-parietal system, the candidate substrate for second-order consciousness and affect BAS and the CBS of primary consciousness. In DAC terms, we can state that the prior values, defined at the level of the RL or CBS, can thus be further coopted and modulated by CL or the fronto-parietal system to follow acquired rules and norms for valuation, mediated by the AL or the amygdala and nucleus accumbens.

(vi) *The definition of qualia*. Given the definition of the H4W problem of survival in the physical world, DAC proposes that fundamental irreducible mental states are motivation (why), knowledge of objects and space (what, where) and time (when). This is a deviation from the classical distinction made by Kant and adapted by Hilgard [166] of knowledge/cognition, feelings/emotions, desires/motivation or connation. DAC predicts that consciousness thus emerged during the Cambrian in response to the recursive and model-based processing required to solve H5W. This was required in order to engage with and adapt to a complex multi-agent and social world pervaded with intentionality. With this, consciousness is a necessary property of embodied and situated control architectures, i.e. brains that engage with their social world of conspecifics, predators, prey and other agents. DAC proposes that sequential unified conscious processing monitors, valuates and modulates parallel real-time control systems both with respect to normativity and the suppression of the intentionality prior. This implies that conscious experience will coalesce around conative and norm relevant states and the context in which they appear. Given that in DAC the agent will become increasingly more dependent on the accumulated experiences in its memory systems from perception, cognition and action to autobiographical and autonoetic states, qualia will also become increasingly more defined by memory and less by sensation. Qualia will express the merging of the embodied memory of the agent with its sensed and imagined reality in the service of the agent's goals.

DACtoc thus builds on the necessary conditions of consciousness captured in the GePe framework and defines consciousness as a virtualization normative memory system (CVNM) that optimizes the parallel control of real-time embodied perception, cognition and action in multi-agent social environments.

(c) Empirical and methodological consequences

If DACtoc is to be a scientific theory, the question is what explanations and testable predictions does it offer? The assumption of the intentionality prior seems to be confirmed by a number of observations. For instance, the perception of biological motion shows that humans ascribe agency to simple moving geometric shapes [167] or moving point models of the body [168]. The intentionality prior hypothesis predicts that infants would ascribe excessive intentionality to the world and that due to maturation and learning this intention reflex is suppressed. Indeed, it has been shown that seven-month-old infants and adults both model the intentional states of others in a form similar to their self-models [169] and that there is a negative correlation between age and the propensity to favour teleological explanations of social behaviour, biological properties, artefacts and life events (e.g. [170]). Given that infants starting at five months of age show slowly maturing neurophysiological signatures of consciousness [171], this opens a wide range of experimental questions on the relation between consciousness, pervasive intentionality and maturation. For instance, how and when is the intentionality reflex suppressed and how is it reflected in the signatures of conscious experience? Here it is important to observe that the prefrontal cortex exerts a strong inhibitory impact on the amygdala and thus its downstream targets. For instance, lesions to the ventromedial prefrontal cortex lead to enhanced activity in the amygdala and associated heightened responses to emotional stimuli [172], while patients suffering from Alzheimer's disease, which can be seen as a disconnection syndrome, show increased arousal and more extreme valence responses to simple sound stimuli [173].

We can further speculate that the intentionality prior underlies religious beliefs, where we can see religion as an anthropomorphized moral believe system. Indeed, one experiment in which theists and atheists had to reason about life events indicated that the former showed a significantly

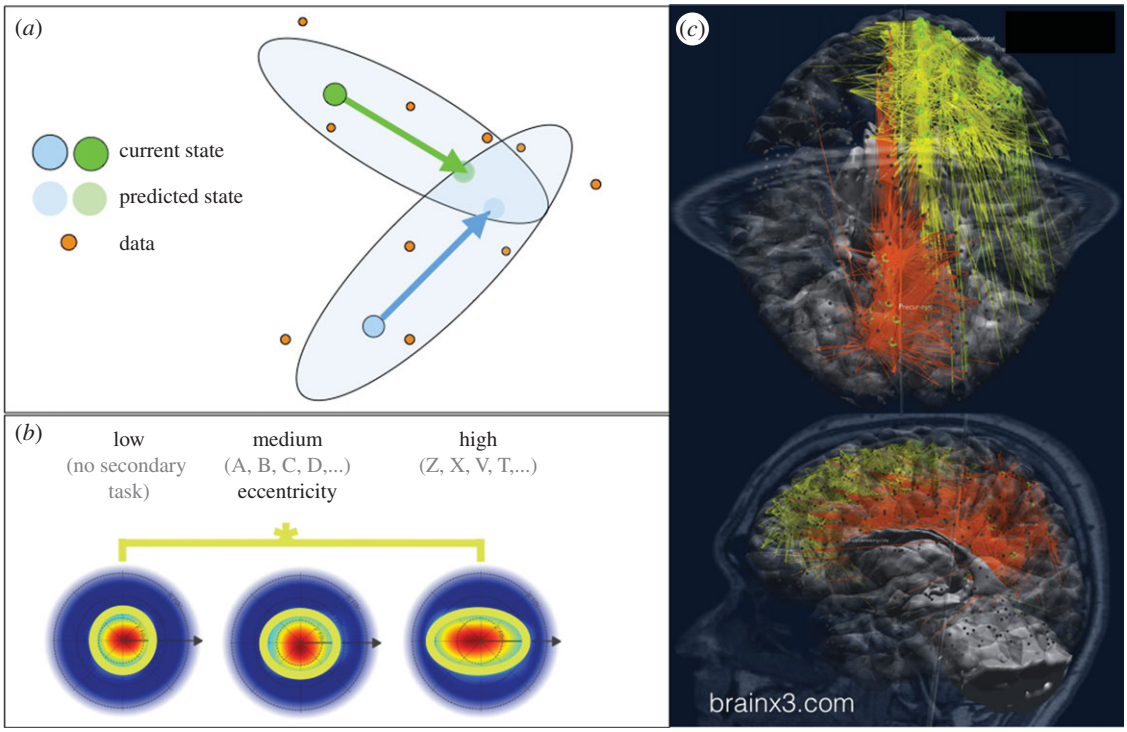

Figure 4. The validation gate hypothesis and the dynamic coupling of consciousness to tasks and sensor states. (*a*) The validation gate hypothesis of perception [178] proposes that information seeking is guided by predictions in a dual form by defining both regions in input space that are expected to provide data and those that do not and thus do not need to be scrutinized. If we follow dots moving along linear trajectories, data (orange dots) is classified relative to areas in input space where it is expected to occur given the properties of stimuli, or their validation gate (light blue area). Resources are only allocated to data which falls outside of the validation gates or when validation gates overlap, i.e. resolving novelty and ambiguity respectively. (*b*) Using a displacement detection task together with reverse correlation allowed us to exactly define the modulation of the validation gates by cognitive load in humans. Increased cognitive load induced a distinct increase of the eccentricity of the validation gate of consciously detected displacements, indicating an expansion of the area that was ignored in sensory processing due to a secondary working memory task (right-hand side kernel). The kernels are displayed as probability distribution of displacements that were followed by a button press from low (left) to high (right) cognitive load. The kernels of fast and slow eye movements did not show a similar effect of cognitive load. Adapted from Mathews *et al*. [179]. (*c*) An fMRI analysis of the validation gate displacement task showed that the explicit detection of displacements were correlated with activation of fronto-parietal networks involving middle and right superior frontal gyrus (Brodmann area 10, 11, yellow), right anterior cingulate cortex (ACC, BA32, yellow) and left precuneus (BA7, orange) (data from Malekshahi *et al*. [180]). These areas are projected onto a three-dimensional reconstruction of the human connectome using brainx3, visualizing the structural and functional connectivity between the nodes (green) of the conscious detection network [181,182].

higher probability to display teleological factors in causal reasoning than the latter, suggesting that atheists are better able to override the intentionality prior [174]. Religious beliefs have been linked to a notion of hyperactive agency detection, which has been interpreted as either an evolutionary epiphenomenon or as serving social processes (see [175] for a review). In addition, an anti-correlation has been identified between inexplicability and the sense of 'awe' and agency detection [176], suggesting that lacking rational explanations, humans resort to the intentionality prior. Interestingly, there is no convincing evidence for a positive correlation between religious beliefs and morality [175], suggesting that it serves an epistemic role in interpreting the world rather than prescribing how to act in it. Hence, the intentionality prior allows us to look at both normativity and religious beliefs from a more concrete operational perspective of how these biases are part and parcel of acting and surviving in an intentional world that we as physical agents have to engage in real time.

Another view of the intentionality prior is to consider the organization of the CBS of the midbrain, or in DAC parlance the brain's reactive control layer. At this level of organization, we find genetically predefined stereotyped action patterns dedicated to the fundamental essential functions sustaining and propagating life, such as security, sex, feeding, grooming and attack, that are triggered by simple initiating stimuli such as the smell of conspecifics, predators, sounds, etc. [32,177]. For instance, freezing to a sudden sound will set in motion the whole machinery behind classical conditioning from CBS and BAS to amygdala, cerebellum and cortex, but above all it is a direct expression of a phylogenetically defined prior: the presence of other agents which intend to consider the observing self as nutrition, triggering the security SEF. In other words, freezing is a behavioural expression of the intentionality prior, threatening intentionality rather than the harmless rustle of leaves.

DACtoc predicts that consciousness is an autonomous virtualization memory system. Hence, its relation to ongoing sensory events must be actively regulated. We have looked directly at this question using goal-oriented psychophysical tasks based on the validation gate hypothesis (figure 4, [178]). The validation gate hypothesis states that perceptual processing exploits predictions to define areas of input space which will and will not receive attentional resources. Negative areas fall within the validation gate, positive areas outside of it (figure 4*a*). We have shown that validation gates for conscious decisions are actively modulated by secondary tasks such that the kernel of the excluded region of input space of a primary detection task expands as the secondary task becomes more demanding (figure 4*b*). This prediction and task-dependent change in the validation

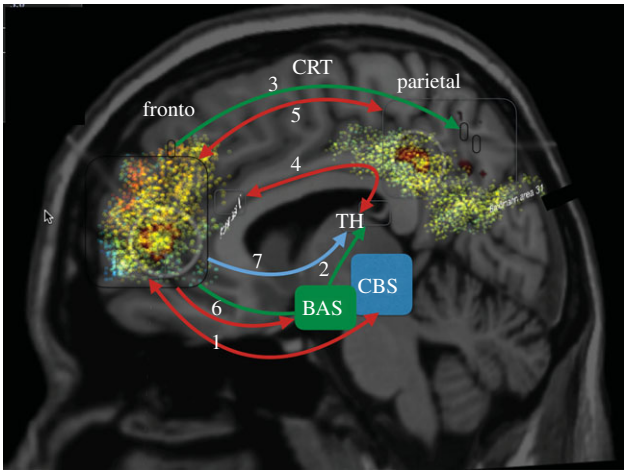

Figure 5. Conceptual diagram of the neuronal substrate of the consciousness autonomous virtualization normative memory system (CVNM). The core behaviour systems (CBS) of the midbrain provide the substrate for primary consciousness driving both the neocortex (1) and the brainstem activating system (BAS) that in turn provide the activation and valuation of the thalamus (2) and neocortex (3). The neocortex (CRT) shares direct and indirect recurrent connections with the thalamus (4) and normative virtualization memory of secondary consciousness is implemented by the cortico-cortical and thalamo-cortical system of the fronto-parietal network (4,5). Cortical networks in turn can exert control over BAS (6) and CBS (1) closing a normative valuation feedback loop. In addition, frontal cortex can inhibit input to sensory thalamic nuclei through projections to the reticular nucleus of the thalamus imposing validation gate driven 'blindness' for task irrelevant states (7). See text for further explanation.

gate has been traced back to activity in core areas of the fronto-parietal network using a fMRI analysis with healthy subjects (figure 4*c*), in particular, the middle and right superior frontal gyrus (Brodmann area 10, 11, yellow), right ACC (BA32, yellow) and left precuneus (BA7, orange) [180]. The latter area is part of the so-called default mode network [183], while the ACC is a central hub of executive control that is hypothesized to exert control over central thalamic nuclei [109], which in turn can switch selected cortical circuits on or off following the general TCD model described above [110,121]. Hence, this fMRI study reveals a circuit for selective prediction-based inhibitory filtering that the validation gate hypothesis proposes, which directly supports the idea that the fronto-parietal system can be considered the substrate of an autonomous consciousness memory system that can selectively couple and uncouple from sensory processing. This latter feature is further invoked by the long-range collaterals that can be found in the thalamo-cortical projections of the prefrontal cortex onto the reticular nucleus of sensory thalamic nuclei facilitating the task specific inhibition of feed-forward sensory processing [184,185].

In contrast with standard models of top-down attention, the validation gate model emphasizes the importance of task-dependent exclusion of sensory information rather than the exclusive amplification of it. As such it does provide an explanation of otherwise surprising attentional phenomena such as change and inattentional blindness (e.g. [186,187]): consciousness memory is actively suppressing input states that are considered irrelevant to the current task given the goals of the agent. In addition, the validation gate experiment showed that conscious decision-making can be dissociated from subconscious parallel processing, providing support for the DAC notion of parallel layered control. Indeed, rapid eye movements, which are controlled by the superior colliculus that is part of the CBS, showed little change across different cognitive load conditions [179]. These experiments also support the idea that consciousness provides a time-delayed description of a virtualized effective task that comprises only a subset of the physical sensory state space the subject is actually acting and existing in. In the fMRI evaluation of the validation gate task, we observed that the medial prefrontal and ACC were engaged during low cognitive load conditions but disengaged from the displacement detection during high cognitive load with a transfer of activity to dorsolateral prefrontal cortex, superior temporal gyrus and the supramarginal gyrus [180]. This reallocation of resources has been observed in other tasks where cognitive load was varied and the stream of consciousness interrupted [188]. In addition, elements in this conscious task memory, such as the orbitofrontal cortex, have been linked to the processing of prediction errors and their valuation by their downstream targets such as the amygdala, VTA and raphe nucleus [189,190]. These observations further support the DACtoc hypothesis that conscious processing is based on an autonomous memory system, which maintains its virtualized intentional state representation independent of external sensory signals, is able to monitor performance and project detected errors to value processing systems.

The validation gate experiment predicts that the CVNM can directly regulate its own input processing by specific inhibition to regions of input space. Recent data shows that this is anatomically and physiologically consistent with the thalamo-cortical system (figure 5). The overall cortico-cortical and thalamo-cortical system implements a counter current architecture where activity can move from posterior to anterior, or from sensory areas towards the CVNM, through direct cortico-cortical and recurrent spiraling cortico-thalamic excitatory projections [191], while CVNM can in turn modulate processing in this system via direct top-down cortico-cortical [192], cortico-thalamic projections [184] and recurrent projections to BAS. In particular the long-range collaterals to the reticular nucleus are of interest with respect to the validation gate hypothesis since these modulate activity of the GABAergic projections to the underlying thalamic nuclei. The validation gate predicts that top-down control over input processing includes the suppression of input by defining regions of no subjective interest rather than solely defining a positive region of interest. Indeed, it has recently been shown that the rodent equivalent of the prefrontal cortex (PFC) directly controls input processing in a task-specific way by driving activity in the sensory zone of the thalamic reticular nucleus [185]. In this way, we can distinguish attention, which in this case acts via an inhibitory cortico-thalamic and an excitatory cortico-cortical pathway, from the access of states to consciousness. The CVNM architecture is part of a thalamo-cortical counter stream processing architecture that can be dynamically switched from operating in a feed-forward sensory-driven to a top-down CVNM-driven mode. We can further speculate that the CVNM is augmented by a thalamo-cortical system that maintains the level of consciousness controlled via the central nucleus of the thalamus [109,193], while emotional context can be controlled through the direct projections from CVNM to the amygdala [184]. Hence, the validation gate hypothesis and its neuronal substrate confirm core principles of the DACtoc hypothesis.

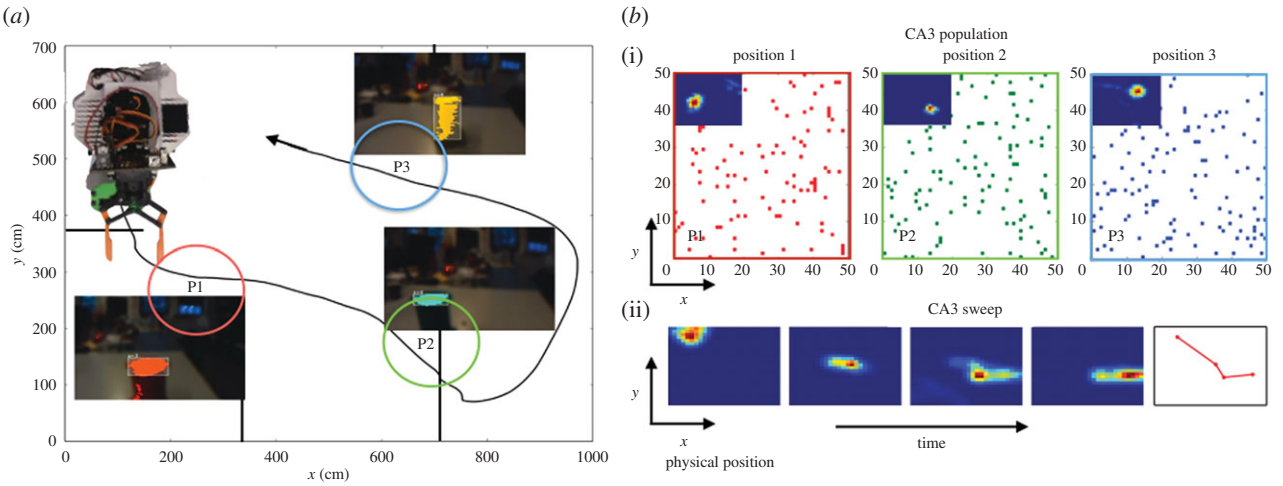

Figure 6. Quale parsing in DACX. (*a*) Example trajectory of a trained DACX agent, showing initial position (left upper corner) and three salient events in the arena where coloured objects are encountered as seen through the robot's camera (locations P1: red, P2: green, P3: blue). (*b*) (i) Population activity in the CA3 episodic memory system of the agent during the three salient navigation events. Colour code as in (*a*). Insets: Rate maps of model single place cells at P1, P2 and P3 showing their spatial specificity. (ii) DACX mind travel by means of a memory sweep showing the goal-oriented engagement of three place cells due to the PFC triggered spreading of activation through the associative connections among the place cells acquired during the exploration of the arena. The actual position of the agent in the upper left corner of the arena is signalled by the first left-most place cell. The drive state was 'hunger' while the goal state was a 'food' item. The right-most panel gives the projected imagined trajectory based on a shortest path search between the activated place cells. Hence, we have parsed a 'hunger' induced, imagined trajectory which leads via two intermediate positions to a location where the agent expects to satisfy its 'food' goal state.

(d) Quale parsing: addressing the 'hard problem' with synthetic consciousness

DACtoc advances a hypothesis on the function, evolutionary origins and neuronal substrate of consciousness. The methodological question is how we can validate it with respect to the content of consciousness. The hard problem is based on the claim that a third person description of first person experience is not possible. I have argued above that this notion is closely tied to a Cartesian reductionist stance on scientific explanation and that this mode of explanation is failing in the face of the multi-scale organization of biological systems. Indeed, one of the critics of Descartes suggested to turn to human experience and skill rather than God to anchor knowledge. The eighteenth-century Neapolitan philosopher Giambattista Vico famously proposed that we can only understand that what we build, or, that 'truth' and 'fact' are reversible (Verum et factum reciprocantur seu convertuntur) [194]. This epistemological model also stood at the heart of the short-lived cybernetics revolution and was directly followed—probably unknowingly—by Watson and Crick in their discovery of the structure of DNA through the construction of a mechanical model. I propose that we can decipher subjective experience following a synthetic strategy [25]. Qualia are defined through the confluence of different modalities of the experienced agent–environment nexus at a specific point in time and given a specific individual history captured in the continuously evolving memory systems of the agent. A third person perspective on qualia will thus require full access to each experiential state over time. This feature is only under very limited experimental control as the nineteenth-century structuralists discovered and experience has thus remained in the first person domain. Full accessibility of all relevant states is only assured through a synthetic approach, where we can control and measure all aspects of an agent starting from the moment it is 'thrown in the world' until the moment it ceases to exist. In his *Monadology*, Leibnitz objected to such a synthetic approach stating that being inside a conscious machine does not reveal its subjective states. Here I propose to solve Leibnitz's challenge by not only entering the machine but above all by having a full record of the history of the system environment interaction that has been captured in its memory systems.

As an example of the quale parsing methodology, we can consider DACX (figure 6). Here, I will focus on DACX's hippocampus because it is the generator of episodic declarative memory providing core content of conscious experience (see [195], for a review), while itself being subconscious [196], i.e. lesions to this area do not affect the level of consciousness but do lead to severe memory impairments. This further confirms the dissociation between processes generating content that can potentially define conscious experience and the transient memory process that serves as the substrate of consciousness itself. The hippocampus shows dense structural and functional integration posing a challenge for integration centric theories of consciousness. In addition, it provides a substrate of virtualization by virtue of its ability to support mental time travel [197–199] through, so-called, sweeps and ripples [200,201], which through spreading activation can re-instantiate and/or readout hippocampal memories that provide information to a range of extra-hippocampal circuits including the default mode network [202].

The entorhinal-hippocampal model of DACX constructs memories of 'episodes' that comprise both spatial and sensory information derived from path integration and the distal sensors the robot is equipped with respectively. These episodic memories are subsequently associated with the goals the agent maintains, facilitating the formation of action policies to optimize performance. Given the ability to follow all relevant state variables of the agent and the environment, each neuronal state can be interpreted in terms of current and/or past experience of the agent. For instance in the example shown, neuronal firing patterns in the CA3 attractor memory system of DACX can be uniquely labelled with respect to the specific location, internal and sensory state of the agent. Most importantly, we can directly interpret the mental time travel the agent engages in. In this case, the agent is searching for the shortest trajectory to a 'food' goal state given an internal need of 'Hunger' (figure 6*b*(ii)). Hence, we have a third person description of the agent's qualia, which

is composed of the experiences of the agent solving a specific instance of the H4W problem and its associated exteroceptive, interoceptive, perceptual, emotional, cognitive and action elements. DACtoc predicts that a similar quale parsing methodology can be applied to conscious synthetic H5W capable agents.

As our models become more elaborate, our ability for quale parsing will also become increasingly more complete, allowing us to unravel the full richness of conscious experience of an artefact. The limitation of the quale parsing method is that it will only be valid when full control over the temporal development of the agent is given, together with full access to its control architecture. This is not feasible with biological systems because we will face a neuroscience version of Heisenberg's principle, where we will in our attempt to measure qualia alter them. A two-step process is required which first maps the biological agent to an artefact and subsequently opens up its experience for third person interpretation. Hence, synthetic quale parsing will not work for arbitrary experiencing systems. However, this is not the scope of the challenge. The problem was to overcome the principled impossibility of accessing first person states. Synthetic quale parsing has answered this challenge, demonstrating the content of the mental time travel of a biologically grounded robot. The 'hard' problem made 'easy'. In order to know what it is like to be a bat, we thus have to emulate this bat and analyse its synthetic alter ego using quale parsing.

5. Discussion

Based on the notion of the H5W problem and the DAC theory of mind and brain, I have argued that consciousness serves the valuation of goal-oriented performance in a world dominated by hidden states in particular derived from the intentionality of other agents and the norms they adhere to. I have argued that in such a world real-time action must be based on parallel automatic control that is optimized with respect to future performance through normative conscious valuation. I have contrasted this proposal with the distracting notion that a hard problem exists in the study of consciousness and the view that quantitative methods provide a theory of this phenomenon. The argument is that these 'solutions' are the result of a misinterpretation of the scientific method: strict reduction will not give us access to the multi-scale organization of mind and brain, neither do methods qualify as theories. For this reason, I proposed an alternative approach which comprises the construction of synthetic consciousness together with a convergent methodology which brings together pertinent constraints from behaviour, anatomy and physiology, in this way, supporting a constructive empiricist view on consciousness where we seek models and theories that are empirically adequate rather than 'true' [203].

The main positions on consciousness can be summarized in one comprehensive framework: GePe. Comparing the components of GePe to the existing DAC architectures, which have been applied to mobile robots, interactive buildings, avatars and humanoids, shows that all components of GePe are present in this architecture, except the puzzling observation that conscious experience trails real-time action. Despite satisfying five out of six GePe criteria, DACX is currently not conscious.

DACtoc proposes that consciousness is critically related to action in an intentional world or the transition from an agent that solves H4W to solving H5W. Here consciousness provides the normative interfacing between the unknowns of the singular embodied self and the social and physical world. In this proposal consciousness is by necessity intentional because it is dealing with a single agent engaged with an intentional world. It is the self-constructed conscious unitary narrative that, grounded in the physical existence of the agent over time, defines its subjectivity and qualia, accumulated in its memory systems and assuring the coherence of the agent's operation. Thus, consciousness is the coherent experience that results from the large-scale integration of parallel processing of perception, affect, memory, cognition and action along the neuroaxis in a dedicated transient memory system that supports unification, virtualization, norm extraction and valuation. It is a form of memory that reflects the autonomous normative states of the agent to facilitate the optimization of its parallel real-time control loops that are driving action and self-regulation. This virtualization normative memory is engaged when the agent plans to act and especially when it evaluates the outcomes of actions in the past, present and future relative to perceived or remembered internal and environmental norms. The core ingredients of consciousness are thus autonomous virtualization memory, intentional simulation-based monitoring and assessment and valuation of parallel multi-scale operations using serialization and unification of goal-oriented performance. The foundations of this process are the intentionality prior and the norm extraction it affords.

Others have also advanced hypotheses that emphasize the social origins of consciousness [61,204–206]. These proposals have rather emphasized the contribution of consciousness to specific aspects of social interaction such as rational thought, language and/or attention. The DACtoc hypothesis emphasizes the role of consciousness in optimizing the control structures that serve action in a world pervaded with hidden states due to the presence of other agents, the social interaction it affords and its underlying intentional stance. Social perception and cognition, however, are part of the subconscious processes driving real-time action that demand a normative solution to credit assignment. Moreover, the self is equally unknowable and pervaded with hidden parallel subconscious states that require a similar treatment of virtualization and serialization. Self, however, will be a less rich source of norms, rather it is subject to them expressing its compliance through the emotions these norms evoke. Hence, DACtoc places norms and the morality they entail at the level of integrated cognitive processes that are appraised through emotions at any level of the DAC hierarchy rather than being defined by them. This is in contrast to and more parsimonious than alternative theories on naturalized morality which see emotions as their grounding [207]. DACtoc makes CVNM dependent on an 'interpreter', which creates a culturally informed autobiographical and normative narrative of a self to which consciousness has only limited direct access consistent with Gazzaniga and Nisbett & Wilson [88,208].

DACtoc is hypothesized to be instantiated in the brain by the combination of CBS, BAS and the thalamo-cortical and cortio-cortical networks of the fronto-parietal system and places its origins in the Cambrian explosion of about 550 Ma. This was the moment that intentionality became a factor in survival. In the treatment of critical transitions in evolution, Smith & Szathmary [209] identify the emergence of sociality as a main transition, paving the way for human societies further aided by the emergence of language. DACtoc is proposing that consciousness was the enabler of this transition. More specifically, that primary consciousness evolved for the initial adjustment to multi-agent environments as in predator–prey systems and

simple eusocial insects, i.e. environments where norms are not hidden but rather unambiguously signalled by the world, while the bootstrapping of secondary consciousness allowed the stabilization of complex social environments as found among mammals yet following rigid prior norms [210]. A key transition omitted by Smith and Szathmary is the evolution of nervous systems and the new forms of information exchange that they afforded [211]. DACtoc proposes that nervous systems facilitated the formation of social groups, in particular, through the emergence of primary and secondary consciousness. Evolutionary progress has been sketched as enhanced autonomy from the environment (see [212] for a review). The DACtoc hypothesis suggests, however, that we might have to consider a tertiary consciousness that is unique to humans, allowing them to redefine their value systems on the basis of socially acquired norms implemented by the feedback control of the neocortex over the norm systems of the CBS and its intentionality prior. This would constitute the ultimate evolutionary transition because it disconnects humans from their phylogenetically defined value systems and the constraints it imposes on their behaviour and propels them into a post-biological Anthropocene: in this sense realizing autonomy not only of the natural environment but also of the biological self and exercising this autonomy by fundamentally changing the world that has given rise to it. We see the effects of this transition daily in ideologically motivated acts of violence and the ecological havoc humans create. DACtoc suggests that CVNM as an adaptive normative system can be trained and motivated to acquire norms and suppress the intentionality prior, creating non-anthropomorphized normative systems. Hence, the post-biological era places even more responsibility on humans to both define moral systems and means of their education, in the service of a sustainable and dignified multi-agent society.

DACtoc reinstates free will as a useful construct to understand the causation of action. DACtoc's CVNM implies that we can and will experience the norms that optimize our parallel action control systems: the ability to *will* improvement of performance in the future as opposed to stopping at the contemporary interpretation that we lack the will to control our performance in the 'now'. Conscious thus only affects future performance and is not the cause of current action. The 'free' aspect of 'free will' can be brought back to the strongly nonlinear properties of a multi-scale brain that operates at the edge of criticality where small changes can make a big difference. It has already been demonstrated in small neuronal models how near criticality can provide for efficient yet system-dependent state space exploration [213]. The question thus becomes what the norms are that we as embodied agents are exposed to how they are embedded in our memories and how they are defining our actions.

It has been common to emphasize levels in the study of consciousness. DACtoc defined consciousness against deficits where both level and content are affected. Indeed, the notion of levels of consciousness has been questioned as being incomplete [92]. DACtoc suggests that the confusion of level and content is due to the fact that they are simultaneously regulated by the CVNM in an anisotropic fashion, i.e. one circuit might be extinguished by becoming dominated by low frequency bursting in its thalamic afferents taking it out of CVNM, while another circuit can be 'ignited' due to its thalamic afferents being released from inhibition. Thus the 'level' of thalamo-cortical circuits are continuously and dynamically re-organized in an anisotropic fashion defining the state of CVNM and content of qualia. The CVNM comprises a key ingredient of the frontal-parietal network that is disrupted in neglect patients [118]. Stroke gives rise to dynamic reorganization of the thalamo-cortical system as expressed in large increases in delta and decreases in alpha activity [121,214]. Taken together this shows that stroke provides an alternative model to investigate consciousness as opposed to sleep or coma, with the advantage to specifically probe the informational aspects of conscious and unconscious processing and their highly variable spatio-temporal organization. In addition, it opens new possibilities to further advance neurorehabilitation and neuroprosthetics methods.

DACtoc advances a network perspective on deficits of consciousness. Indeed, the fronto-parietal network is centrally connected to all identified functional networks of the neocortex [215]. Disturbances to this network should lead to pathologies in norm-dependent behaviour. DACtoc predicts that borderline personality disorder, a pathology with emotional, cognitive and behavioural dysregulation [216], is largely due to instabilities in the feedback between systems underlying normative secondary and valuative primary consciousness and the autobiographical narrative of the self. Indeed, these patients show strongly different responses to norm violations as compared to control subjects for instance showing an enhanced tendency to display 'angry retaliation' [217]. In addition, we can speculate that distortions of conscious experience that occur during psychosis are due to the dysregulation of the coupling between CVNM and sensory processes following the DACtoc-associated validation gate hypothesis, as schizophrenia is also associated with dysfunction of the thalamic reticular nucleus [218]. These predictions are currently being pursued in various clinical settings.

Recently, an argument has been made that the 'hot spot' for content consciousness can be found in the posterior cortex rather than the fronto-parietal network and/or the thalamo-cortical system [4]. This suggestion, however, might have been biased by the highly selective emphasis on non-reported sensory processing alone. The example of neglect used here to define consciousness also illustrates that the content of experience depends on the modalities considered. These include executive function and action such as expressed in the goals being pursued, affect and autobiographical memory, defining the foundations of autonoetic experience. Neglect is a multi-faceted pathology due to a distributed network destabilization of the thalamo-cortical system; the same holds for other deficits of consciousness including schizophrenia [219]. Hence, there is no convincing evidence to depart from the multi-scale network interpretation of consciousness advocated here.

DACtoc proposes that consciousness is implemented by means of a two-step mechanism involving BAS and midbrain CBS, initially serving primary consciousness and later coopted by the thalamus and the fronto-parietal system of the neocortex in the virtualizing normative transient memory of secondary consciousness. One aspect of this hypothesis is that none of the systems attributed to the AL of DAC, in particular, the hippocampus, amygdala, cerebellum and basal ganglia, are figuring in the realization of primary and secondary consciousness. A common feature of these systems is that they are involved in the identification of the agent's state space using extero-, intero- and proprioceptive signals combined with massive parallel processing we can speculate that this provides the core subconscious and probabilistic set of states that are critical in the generation of parallel real-time control on which CVNM operates. Indeed, the cerebellum is the example par excellence

of an adaptive real-time controller, which can provide high-resolution event triggers in a time window of about 1 s in classical conditioning [220]. DACtoc proposes that this feature of the AL is an anatomical signature of large-scale unconscious parallel real-time processing that in turn requires CVNM, found in social brains. Another question with respect to the substrate of consciousness is how the stream of consciousness is updated and its content controlled. Given its cortical dynamics, we can take inspiration from the update dynamics of other cortical systems. Here at least two predictions stand out. Firstly, it has been proposed that the striatonigrostriatal pathway sets up a specific sequential processing stream in the dynamics of the basal ganglia, forming an ascending spiral that defines a hierarchy of information flow from the ventral striatum's shell and core (motivation) to central (cognition) and dorsolateral striatum (action) [221]. The striatum is the cortical input stage of the basal ganglia, which is providing selection of cortical states via its projections to the thalamus, as we saw previously, and we can speculate that this subcortical sequencer will also dictate the update frequency of the CVNM of the fronto-parietal system and the selection of its content. We can speculate that a similar process underlies the integration and synchronization of the parietal and frontal zones of the CVNM. Secondly, the onset of movement is accompanied in the primary motor cortex by the attenuation of beta activity [222]. This suggests that beta activity prevents readout by M1 neurons of the motor plans formed in premotor. The attenuation of beta activity allows to read in a new motor program, which can be played out towards the periphery. This sequential process is consistent with the basal ganglia sequencer notion and suggests that given the similarity in computational hardware, the updates of CVNM will follow a comparable process with the difference that in between updates, information must be held in transient memory of the thalamo-cortical loops rather than being silenced by beta activity. An associated question is how conscious content can be flexibly refreshed and or maintained in a transient memory system. Popular notions of rate coding [223] would require a labelled line hierarchical coding system, where each synapse and neuron in the system has a specific semantic dedication. How can such a solution provide a representational substrate for the practically infinite set of phenomenal states, i.e. differentiation? The popular labelled line approach, although easy in digital hardware, confronts biological control systems with insurmountable problems of packing sufficient wires and neurons in strategic functionally defined locations. Moreover, fronto-parietal networks appear to provide more general-purpose hardware, which can be flexibly dedicated to widely varying tasks [215,224]. Hence, this leads to the prediction that the CVNM relies on temporally encoded information to facilitate rapid information refresh and multiplexing [225].

The DACtoc explanation of consciousness focuses on the unitary nature of conscious experience, its normative role and its delayed realization relative to real-time performance. The model behind this proposal has been advanced as an embodied robot-based system following the idea that empirically adequate models of embodied and situated brains will be machines themselves. For this reason, we have embarked on a series of DAC control architectures for humanoid robots that in particular advance human-level embodied social interaction [226]. In order to realize its convergent validation, we have also integrated this humanoid platform with state-of-the-art human connectome models using brainx3 [181,182] that are augmented with fully simulated subcortical structures. DACtoc thus predicts that human-level quale parsing will result from the convergence of humanoid robotics and whole brain modelling following the convergent validation methodology.

There is quite some interest in the field of machine consciousness where different aspects described in the GePe framework are addressed (see for a recent review [227], this review is surprisingly bleak about the outlook for this field though). An important difference between DACtoc and other approaches is that DACtoc is building on a long tradition of biologically grounded brain models, is advanced in embodied and situated form and considers consciousness as a necessary ingredient of survival in a social world.

I have shown that a synthetic constructive methodology allows us to parse qualia in terms of the complex mental time travel of a biologically grounded embodied control architecture, by providing full control over its development over time and the internal states of its subsystems. Following this approach, we can have a third person perspective on 'what it is like' to be a bat or even a philosopher. The implication is, however, that we do have to construct an embodied situated real-world model of the agent in question. This raises the important question of the granularity of our answer to the hard problem. Do we really need to explain, predict and control every individual quale or rather provide insight in the processes that shape them? The latter option seems closer to what we can expect from the scientific method and need for practical applications, the former version of HPEP is great for science fiction novels.

I have contrasted the DACtoc methodology with the hard problem and IIT and benchmarked it against the dominant trends in theories of consciousness captured in GePe. I have identified as the missing piece in the puzzle of consciousness, its function of extracting norms from the hidden states of the social world in order to optimize parallel real-time action control. I have argued that the trend to turn away from the questions on the ontology and function of consciousness is dissatisfying from an intellectual perspective. More importantly, it also discharges science from its responsibility towards building a sustainable and dignified society. If science is supposed to provide explanations, predictions and control of natural phenomena then science's success should also be measured in terms of its impact. It should not only be able to contribute to pressing challenges in the domains of education, health and well-being but especially due to the secular turn in modern Western societies, provide a foundation for the grounding of our metaphysics. Answering the question of what consciousness is and how physical systems can give rise to it, stands at the centre of knowing what it is to be human and to face up to the fundamental challenges of our time and any time in which conscious beings have existed and will exist in the future. We must and will be gnawing this file forever!

Competing interests. I declare I have no competing interests.

Funding. The work reported here is supported by EC FP7 project Convergent Science Network (CSN FP7-ICT-601167), European Research Council grant cDAC (ERC-341196) and EC FP7 project WYSIWYD (FP7-612139).

Acknowledgements. I thank the members of SPECS for their feedback on this manuscript. In particular, Riccardo Zucca, Xerxes Arsiwalla, Martina Mayer, Greg Zegarek and Sytse Wierenga and David Dalmazzo for assistance in realizing the artwork. I have profited from discussions with many colleagues, in particular, Bjorn Merker, Stephen Deiss and Anil Seth. In addition, I thank two of the three reviewers of this article for their very valuable and constructive feedback.

References

1. Crick F, Koch C. 1990 Towards a neurobiological theory of consciousness. *Semin. Neurosci.* **2**, 263–275. (doi:10.1016/B978-0-12-185254-2.50021-8)
2. Edelman GM. 1989 *The remembered present: a biological theory of consciousness*. New York, NY: Basic Books.
3. Dehaene S, Changeux JP. 2011 Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227. (doi:10.1016/j.neuron.2011.03.018)
4. Koch C *et al.* 2016 Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* **17**, 307–321. (doi:10.1038/nrn.2016.22)
5. Finkelstein G. 2014 Emil du Bois-Reymond's reflections on consciousness. In *Brain, mind and consciousness in the history of neuroscience* (eds CUM Smith, H Whitaker), pp. 163–184. Dordrecht, The Netherlands: Springer.
6. Chalmers D. 1995 Facing up to the problem of consciousness. *J. Conscious. Stud.* **2**, 200–219.
7. Levine J. 1983 Materialism and qualia: the explanatory gap. *Pac. Philos. Q.* **64**, 354–361.
8. Nagel T. 1974 What is it like to be a bat. *Philos. Rev.* **83**, 435–450. (doi:10.2307/2183914)
9. Block N. 1995 How many concepts of consciousness? *Behav. Brain Sci.* **18**, 272–287. (doi:10.1017/S0140525X00038486)
10. de Spinoza B. 1677/1922 Ethica. (Transl. by W Meijer). Amsterdam, The Netherlands: SL. van Looy.
11. Verschure PFMJ. 2012 Distributed adaptive control: a theory of the mind, brain, body nexus. *Biol. Inspired Cognit. Archit.* **1**, 55–72. (doi:10.1016/j.bica.2012.04.005)
12. Verschure PFMJ, Voegtlin T, Douglas RJ. 2003 Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature* **425**, 620–624. (doi:10.1038/nature02024)
13. Woese CR. 2004 A new biology for a new century. *Microbiol. Mol. Biol. Rev.* **68**, 173–186. (doi:10.1128/MMBR.68.2.173-186.2004)
14. Goldberg AD, Allis CD, Bernstein E. 2007 Epigenetics: a landscape takes shape. *Cell* **128**, 635–638. (doi:10.1016/j.cell.2007.02.006)
15. Kirschner MW, Gerhart JC. 2006 *The plausibility of life: resolving Darwin's dilemma*. New Haven, CT: Yale University Press.
16. Mychasiuk R, Gibb R, Kolb B. 2012 Prenatal stress alters dendritic morphology and synaptic connectivity in the prefrontal cortex and hippocampus of developing offspring. *Synapse* **66**, 308–314. (doi:10.1002/syn.21512)
17. Chalmers DJ. 2010 *The character of consciousness*. Oxford, UK: Oxford University Press.
18. Tononi G, Koch C. 2015 Consciousness: here, there and everywhere? *Phil. Trans. R. Soc. B* **370**, 20140167. (doi:10.1098/rstb.2014.0167)
19. Verschure PFMJ. In press. From Big Data back to Big Ideas: the risks of a theory free data rich science of mind and brain and a solution. *Connect. Sci.*
20. Kline M. 1985 *Mathematics and the search for knowledge*. Oxford, UK: Oxford University Press.
21. Price H. 2002 Boltzmann's time bomb. *Br. J. Philos. Sci.* **53**, 83–119. (doi:10.1093/bjps/53.1.83)
22. James W. 1950 The principles of psychology (1890, vol. 1). New York, NY: Holt.
23. Deacon TW. 2011 *Incomplete nature: how mind emerged from matter*. New York, NY: WW Norton.
24. Dobzhansky T. 1973 Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **35**, 125–129. (doi:10.2307/4444260)
25. Verschure PFMJ. 1997 Connectionist explanation: taking positions in the mind-brain dilemma. In *Neural networks and a new artificial intelligence* (ed. G Dorffner), pp. 133–188. London, UK: Thompson.
26. Verschure PFMJ. 2016 Consciousness in action: the unconscious parallel present optimized by the conscious sequential projected future. In *Where's the action?: the pragmatic turn in cognitive science* (eds AK Engel, K Frniston, D Kragic). Boston, MA: MIT Press.
27. Verschure PFMJ, Pennartz CMA, Pezzulo G. 2014 The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Phil. Trans. R. Soc. B* **369**, 20130479. (doi:10.1098/rstb.2013.0483)
28. Sanchez-Fibla M *et al.* 2010 Allostatic control for robot behavior regulation: a comparative rodent-robot study. *Adv. Complex Syst.* **13**, 377–403. (doi:10.1142/S0219525910002621)
29. Verschure PFMJ, Pfeifer R. 1992 Categorization, representations, and the dynamics of system-environment interaction: a case study in autonomous systems. In *From animals to animats: proceedings of the second international conference on simulation of adaptive behavior*. Cambridge, MA: MIT Press.
30. Duff A, Verschure PF. 2010 Unifying perceptual and behavioral learning with a correlative subspace learning rule. *Neurocomputing* **73**, 1818–1830. (doi:10.1016/j.neucom.2009.11.048)
31. Merker B. 2013 The efference cascade, consciousness, and its self: naturalizing the first person pivot of action control. *Front. Psychol.* **4**, 501. (doi:10.3389/fpsyg.2013.00501)
32. Panksepp J, Biven L. 2012 *The archaeology of mind: neuroevolutionary origins of human emotions* (Norton series on interpersonal neurobiology). New York, NY: WW Norton & Company.
33. Verschure PFMJ, Kröse BJA, Pfeifer R. 1992 Distributed adaptive control: the self-organization of structured behavior. *Robot. Auton. Syst.* **9**, 181–196. (doi:10.1016/0921-8890(92)90054-3)
34. Verschure PFMJ *et al.* 1995 Multilevel analysis of classical conditioning in a behaving real world artifact. *Robot. Auton. Syst.* **16**, 247–265. (doi:10.1016/0921-8890(95)00050-X)
35. Maffei G *et al.* 2015 An embodied biologically constrained model of foraging: from classical and operant conditioning to adaptive real-world behavior in DAC-X. *Neural Netw.* **72**, 88–108. (doi:10.1016/j.neunet.2015.10.004)
36. Metzinger T. 2003 *Being no one: the self-model theory of subjectivity*. Cambridge, MA: MIT Press.
37. Frith CD. 2008 Social cognition. *Phil. Trans. R. Soc. B* **363**, 2033–2039. (doi:10.1098/rstb.2008.0005)
38. Verschure PFMJ. 1992 Taking connectionism seriously: the vague promise of subsymbolism and an alternative. In *Proc. of the 14th Annu. Conf. of the Cognitive Science Society*, pp. 653–658. Hillsdale, NJ: Erlbaum.
39. Pfeifer R, Bongard JC. 2006 *How the body shapes the way we think: a new view of intelligence*. Cambridge, MA: MIT Press.
40. Damasio A. 2012 *Self comes to mind: constructing the conscious brain*. New York, NY: Random House.
41. Morsella E *et al.* 2015 Homing in on consciousness in the nervous system: an action-based synthesis. *Behav. Brain Sci.* First View, 1–106. (doi:10.1017/S0140525X15000643)
42. Brugger P *et al.* 2000 Beyond re-membering: phantom sensations of congenitally absent limbs. *Proc. Natl Acad. Sci. USA* **97**, 6167–6172. (doi:10.1073/pnas.100510697)
43. Pavlov IP. 1927 *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. London, UK: Oxford University Press.
44. Varela FJ, Thompson ET, Rosch E. 1992 *The embodied mind: cognitive science and human experience*. Cambridge, MA: MIT Press.
45. O'Regan JK, Noe A. 2001 A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* **24**, 939–973; discussion 973–1031. (doi:10.1017/S0140525X01000115)
46. von Helmholtz H. 1924 *Helmholtz's treatise on physiological optics*, vols 1–3. Chelmsford, MA: Courier Corporation.
47. Tolman EC. 1932 *Purposive behavior in animals and man*. New York, NY: Century Co.
48. Craik KJW. 1943 *The nature of explanation*. Cambridge, UK: Cambridge University Press.
49. Bar M. 2007 The proactive brain: using analogies and associations to generate predictions. *Trends Cognit. Sci.* **11**, 280–289. (doi:10.1016/j.tics.2007.05.005)
50. Hesslow G. 2002 Conscious thought as simulation of behaviour and perception. *Trends Cognit. Sci.* **6**, 242–247. (doi:10.1016/S1364-6613(02)01913-7)
51. Barsalou LW. 2008 Grounded cognition. *Annu. Rev. Psychol.* **59**, 617–645. (doi:10.1146/annurev.psych.59.103006.093639)
52. Revonsuo A. 2006 *Inner presence: consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.
53. Merker B. 2005 The liabilities of mobility: a selection pressure for the transition to consciousness in animal evolution. *Conscious. Cogn.* **14**, 89–114. (doi:10.1016/S1053-8100(03)00002-3)
54. Massaro DW. 1997 *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
55. Friston K. 2010 The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138. (doi:10.1038/nrn2787)

56. Clark A. 2013 Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204. (doi:10.1017/S0140525X12000477)
57. Herreros I, Verschure PF. 2015 About the goal of a goals' goal theory. *Cognit. Neurosci.* **6**, 218–219. (doi:10.1080/17588928.2015.1051952)
58. Boly M *et al.* 2011 Preserved feedforward but impaired top-down processes in the vegetative state. *Science* **332**, 858–862. (doi:10.1126/science.1202043)
59. Bekinschtein TA *et al.* 2009 Classical conditioning in the vegetative and minimally conscious state. *Nat. Neurosci.* **12**, 1343–1349. (doi:10.1038/nn.2391)
60. Ward LM. 2011 The thalamic dynamic core theory of conscious experience. *Conscious. Cognit.* **20**, 464–486. (doi:10.1016/j.concog.2011.01.007)
61. Graziano MS. 2013 *Consciousness and the social brain*. Oxford, UK: Oxford University Press.
62. Tononi G, Edelman GM. 1998 Consciousness and complexity. *Science* **282**, 1846–1851. (doi:10.1126/science.282.5395.1846)
63. Arsiwalla XD, Verschure PF. 2013 Integrated information for large complex networks. In *The IEEE Int. Joint Conf. on Neural Networks* (*IJCNN*). New York, NY: IEEE.
64. Arsiwalla XD, Verschure PFMJ. 2016 Computing information integration in brain networks. In *Advances in network science*. Berlin, Germany: Springer International Publishing.
65. Rosenthal DM. 2008 Consciousness and its function. *Neuropsychologia* **46**, 829–840. (doi:10.1016/j.neuropsychologia.2007.11.012)
66. Cerullo MA. 2015 The problem with phi: a critique of integrated information theory. *PLoS Comput. Biol.* **11**, e1004286. (doi:10.1371/journal.pcbi.1004286)
67. Koch C, Tononi G. 2008 Can machines be conscious? *IEEE Spectrum* **45**, 55–59. (doi:10.1109/MSPEC.2008.4531463)
68. Baars BJ. 1988 *A cognitive theory of consciousness*. New York, NY: Cambridge University Press.
69. Block N. 2007 Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* **30**, 481–499; discussion 499–548. (doi:10.1017/S0140525X07002786)
70. Dehaene S. 2014 *Consciousness and the brain: deciphering how the brain codes our thoughts*. Harmondsworth, UK: Penguin.
71. Wegner DM. 2003 *The illusion of conscious will*. Cambridge, MA: MIT Press.
72. Dennett DC. 1992 *Consciousness explained*. New York, NY: Little Brown.
73. Milner D, Goodale M. 1995 *The visual brain in action*. Oxford, UK: Oxford University Press.
74. Mark RN *et al.* 2015 On making the right choice: a meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment Decis. Making* **10**, 1–17.
75. Dijksterhuis A, Bargh JA. 2001 The perception-behavior expressway: automatic effects of social perception on social behavior. *Adv. Exp. Soc. Psychol.* **33**, 1–40. (doi:10.1016/S0065-2601(01)80003-4)
76. Frith CD, Blakemore SJ, Wolpert DM. 2000 Abnormalities in the awareness and control of action. *Phil. Trans. R. Soc. Lond. B* **355**, 1771. (doi:10.1098/rstb.2000.0734)
77. Desmurget M, Sirigu A. 2009 A parietal-premotor network for movement intention and motor awareness. *Trends Cognit. Sci.* **13**, 411–419. (doi:10.1016/j.tics.2009.08.001)
78. Libet B. 1985 Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behav. Brain Sci.* **8**, 529–566. (doi:10.1017/S0140525X00044903)
79. Haggard P. 2008 Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* **9**, 934–946. (doi:10.1038/nrn2497)
80. Custers R, Aarts H. 2010 The unconscious will: how the pursuit of goals operates outside of conscious awareness. *Science* **329**, 47–50. (doi:10.1126/science.1188595)
81. Tsuchiya N, Adolphs R. 2007 Emotion and consciousness. *Trends Cogn. Sci.* **11**, 158–167. (doi:10.1016/j.tics.2007.01.005)
82. Critchley HD *et al.* 2004 Neural systems supporting interoceptive awareness. *Nat. Neurosci.* **7**, 189–195. (doi:10.1038/nn1176)
83. Uhlhaas PJ *et al.* 2009 Neural synchrony in cortical networks: history, concept and current status. *Front. Integr. Neurosci.* **3**, 17. (doi:10.3389/neuro.07.017.2009)
84. Dijksterhuis A, Aarts H. 2010 Goals, attention, and (un) consciousness. *Annu. Rev. Psychol.* **61**, 467–490. (doi:10.1146/annurev.psych.093008.100445)
85. Baumeister RF, Masicampo E, Vohs KD. 2011 Do conscious thoughts cause behavior? *Annu. Rev. Psychol.* **62**, 331–361. (doi:10.1146/annurev.psych.093008.131126)
86. Haggard P, Eimer M. 1999 On the relation between brain potentials and the awareness of voluntary movements. *Exp. Brain Res.* **126**, 128–133. (doi:10.1007/s002210050722)
87. Kahneman D. 2011 *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
88. Gazzaniga MS. 2011 *Who's in charge?: free will and the science of the brain*. New York, NY: Harper Collins.
89. Evans JSBT. 2008 Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–278. (doi:10.1146/annurev.psych.59.103006.093629)
90. Schurger A, Sitt JD, Dehaene S. 2012 An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proc. Natl Acad. Sci. USA* **109**, E2904–E2913. (doi:10.1073/pnas.1210467109)
91. Di Perri C *et al.* 2014 Functional neuroanatomy of disorders of consciousness. *Epilepsy Behav.* **30**, 28–32. (doi:10.1016/j.yebeh.2013.09.014)
92. Bayne T, Hohwy J, Owen AM. 2016 Are there levels of consciousness? *Trends Cognit. Sci.* **20**, 405–413. (doi:10.1016/j.tics.2016.03.009)
93. Merker B. 2007 Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav. Brain Sci.* **30**, 63–81. (doi:10.1017/S0140525X07000891)
94. Penfield W, Jasper H. 1954 *Epilepsy and the functional anatomy of the human brain*. Boston, MA: Little Brown.
95. Panksepp J. 2008 The affective brain and core consciousness: how does neural activity generate emotional feelings? In *Handbook of emotions*, (eds M Lewis, JM Haviland-Jones, LF Barrett), 3rd edn pp. 47–67. New York, NY: Guilford Press.
96. Parvizi J, Damasio A. 2001 Consciousness and the brainstem. *Cognition* **79**, 135–160. (doi:10.1016/S0010-0277(00)00127-X)
97. Barron AB, Klein C. 2016 What insects can tell us about the origins of consciousness. *Proc. Natl Acad. Sci. USA* **113**, 4900–4908. (doi:10.1073/pnas.1520084113)
98. Strausfeld NJ, Hirth F. 2013 Deep homology of arthropod central complex and vertebrate basal ganglia. *Science* **340**, 157–161. (doi:10.1126/science.1231828)
99. Feinberg TE, Mallat J. 2013 The evolutionary and genetic origins of consciousness in the Cambrian period over 500 million years ago. *Front. Psychol.* **4**, 1–27. (doi:10.3389/fpsyg.2013.00667)
100. Mashour GA, Alkire MT. 2013 Evolution of consciousness: phylogeny, ontogeny, and emergence from general anesthesia. *Proc. Natl Acad. Sci. USA* **110**(Suppl 2), 10 357–10 364. (doi:10.1073/pnas.1301188110)
101. Gandhi NJ, Katnani HA. 2011 Motor functions of the superior colliculus. *Annu. Rev. Neurosci.* **34**, 205–231. (doi:10.1146/annurev-neuro-061010-113728)
102. Långsjö JW *et al.* 2012 Returning from oblivion: imaging the neural core of consciousness. *J. Neurosci.* **32**, 4935–4943. (doi:10.1523/JNEUROSCI.4962-11.2012)
103. Parvizi J, Damasio AR. 2003 Neuroanatomical correlates of brainstem coma. *Brain* **126**, 1524–1536. (doi:10.1093/brain/awg166)
104. Jones BE. 2003 Arousal systems. *Front. Biosci.* **8**, s438–s451. (doi:10.2741/1074)
105. Owen AM *et al.* 2006 Detecting awareness in the vegetative state. *Science* **313**, 1402. (doi:10.1126/science.1130197)
106. Schiff ND *et al.* 2007 Behavioural improvements with thalamic stimulation after severe traumatic brain injury. *Nature* **448**, 600–603. (doi:10.1038/nature06041)
107. Steriade M. 2005 Sleep, epilepsy and thalamic reticular inhibitory neurons. *Trends Neurosci.* **28**, 317–324. (doi:10.1016/j.tins.2005.03.007)
108. Llinás RR *et al.* 1999 Thalamocortical dysrhythmia: a neurological and neuropsychiatric syndrome characterized by magnetoencephalography. *Proc. Natl Acad. Sci. USA* **96**, 15 222–15 227. (doi:10.1073/pnas.96.26.15222)
109. Schiff ND. 2010 Recovery of consciousness after brain injury: a mesocircuit hypothesis. *Trends Neurosci.* **33**, 1–9. (doi:10.1016/j.tins.2009.11.002)
110. Proske H, Jeanmonod D, Verschure PFMJ. 2011 A computational model of thalamocortical dysrhythmia. *Eur. J. Neurosci.* **33**, 1281–1290. (doi:10.1111/j.1460-9568.2010.07588.x)
111. Adams JH, Graham D, Jennett B. 2000 The neuropathology of the vegetative state after an acute brain insult. *Brain* **123**, 1327–1338. (doi:10.1093/brain/123.7.1327)

112. Grubb BP *et al.* 1998 Cerebral syncope: loss of consciousness associated with cerebral vasoconstriction in the absence of systemic hypotension. *Pacing Clin. Electrophysiol.* **21**, 652–658. (doi:10.1111/j.1540-8159.1998.tb00120.x)
113. Blumenfeld H. 2012 Impaired consciousness in epilepsy. *Lancet Neurol.* **11**, 814–826. (doi:10.1016/S1474-4422(12)70188-6)
114. Englot DJ *et al.* 2010 Impaired consciousness in temporal lobe seizures: role of cortical slow activity. *Brain* **133**, 3764–3777. (doi:10.1093/brain/awq316)
115. Corbetta M. 2014 Hemispatial neglect: clinic, pathogenesis, and treatment. *Semin. Neurol.* **34**, 514–523. (doi:10.1055/s-0034-1396005)
116. Parton A, Malhotra P, Husain M. 2004 Hemispatial neglect. *J. Neurol. Neurosurg. Psychiatry* **75**, 13–21.
117. Ptak R. 2012 The frontoparietal attention network of the human brain: action, saliency, and a priority map of the environment. *Neuroscientist* **18**, 502–515. (doi:10.1177/1073858411409051)
118. De Schotten MT *et al.* 2014 Damage to white matter pathways in subacute and chronic spatial neglect: a group study and 2 single-case studies with complete virtual '*in vivo*' tractography dissection. *Cereb. Cortex* **24**, 691–706. (doi:10.1093/cercor/bhs351)
119. Fornito A, Zalesky A, Breakspear M. 2015 The connectomics of brain disorders. *Nat. Rev. Neurosci.* **16**, 159–172. (doi:10.1038/nrn3901)
120. Dubovik S *et al.* 2012 The behavioral significance of coherent resting-state oscillations after stroke. *Neuroimage* **61**, 249–257. (doi:10.1016/j.neuroimage.2012.03.024)
121. van Wijngaarden JBG *et al.* In press. The development of thalamo-cortical dysrhythmia after acute ischaemic stroke: a combined experimental and theoretical study. *Public Libr. Sci. Comput. Biol.*
122. Orfei M *et al.* 2007 Anosognosia for hemiplegia after stroke is a multifaceted phenomenon: a systematic review of the literature. *Brain* **130**, 3075–3090. (doi:10.1093/brain/awm106)
123. Garbarini F *et al.* 2012 'Moving' a paralysed hand: bimanual coupling effect in patients with anosognosia for hemiplegia. *Brain* **135**, 1486–1497. (doi:10.1093/brain/aws015)
124. Campillo C *et al.* 2011 The visual processing of local and global features in stroke patients. Paper presented at the *20th Eur. Stroke Conf., 24–27 May, Hamburg, Germany*.
125. Mijovic-Prelec D *et al.* 1998 The judgement of absence in neglect. *Neuropsychologia* **36**, 797–802. (doi:10.1016/S0028-3932(97)00144-9)
126. Bayne T. 2010 *The unity of consciousness*. Oxford, UK: Oxford University Press.
127. Mashour GA. 2006 Integrating the science of consciousness and anesthesia. *Anesth. Analg.* **103**, 975–982. (doi:10.1213/01.ane.0000232442.69757.4a)
128. Weinberger NM. 2004 Experience-dependent response plasticity in the auditory cortex: issues, characteristics, mechanisms and functions. In *Plasticity of the auditory system* (eds TN Parks, EW Rubel, RR Fay), pp. 173–228. New York, NY: Springer.
129. Van Ruitenbeek P, Vermeeren A, Riedel W. 2010 Cognitive domains affected by histamine H1-antagonism in humans: a literature review. *Brain Res. Rev.* **64**, 263–282. (doi:10.1016/j.brainresrev.2010.04.008)
130. den Ouden HE *et al.* 2013 Dissociable effects of dopamine and serotonin on reversal learning. *Neuron* **80**, 1090–1100. (doi:10.1016/j.neuron.2013.08.030)
131. Sara SJ. 2009 The locus coeruleus and noradrenergic modulation of cognition. *Nat. Rev. Neurosci.* **10**, 211–223. (doi:10.1038/nrn2573)
132. LeDoux J. 2012 Rethinking the emotional brain. *Neuron* **73**, 653–676. (doi:10.1016/j.neuron.2012.02.004)
133. Selvaraj S *et al.* 2014 Alterations in the serotonin system in schizophrenia: a systematic review and meta-analysis of postmortem and molecular imaging studies. *Neurosci. Biobehav. Rev.* **45**, 233–245. (doi:10.1016/j.neubiorev.2014.06.005)
134. Rennó-Costa C, Lisman JE, Verschure PFMJ. 2010 The mechanism of rate remapping in the dentate gyrus. *Neuron* **68**, 1051–1058. (doi:10.1016/j.neuron.2010.11.024)
135. Lu L *et al.* 2013 Impaired hippocampal rate coding after lesions of the lateral entorhinal cortex. *Nat. Neurosci.* **16**, 1085–1093. (doi:10.1038/nn.3462)
136. Verschure PFMJ, Althaus P. 2003 A real-world rational agent: unifying old and new AI. *Cognit. Sci.* **27**, 561–590. (doi:10.1207/s15516709cog2704_1)
137. Moutard C, Dehaene S, Malach R. 2015 Spontaneous fluctuations and non-linear ignitions: two dynamic faces of cortical recurrent loops. *Neuron* **88**, 194–206. (doi:10.1016/j.neuron.2015.09.018)
138. Premack D, Woodruff G. 1978 Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**, 515–526. (doi:10.1017/S0140525X00076512)
139. Tomasello M *et al.* 2005 Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* **28**, 675–691; discussion 691–735. (doi:10.1017/S0140525X05000129)
140. Inderbitzin M *et al.* 2013 The social perceptual salience effect. *J. Exp. Psychol.* **39**, 62–74. (doi:10.1037/a0028317)
141. Herculano-Houzel S. 2009 The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum Neurosci.* **3**, 1–11. (doi:10.3389/neuro.09.031.2009)
142. Verschuuren J *et al.* 1996 Inflammatory infiltrates and complete absence of Purkinje cells in anti-Yo-associated paraneoplastic cerebellar degeneration. *Acta Neuropathol.* **91**, 519–525. (doi:10.1007/s004010050460)
143. Fiorio M *et al.* 2014 The role of the cerebellum in dynamic changes of the sense of body ownership: a study in patients with cerebellar degeneration. *J. Cognit. Neurosci.* **26**, 712–721. (doi:10.1162/jocn_a_00522)
144. Buckner RL. 2013 The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. *Neuron* **80**, 807–815. (doi:10.1016/j.neuron.2013.10.044)
145. Caligiore D *et al.* In press. Consensus paper: towards a systems-level view of cerebellar function: the interplay between cerebellum, basal ganglia, and cortex. *Cerebellum*. (doi:10.1007/s12311-016-0763-3)
146. Rochat P. 2009 *Others in mind: social origins of self-consciousness*. Cambridge, UK: Cambridge University Press.
147. Verschure PFMJ. 2012 Consciousness solves pervasive intentionality. In *The 16th Annu. Meet. Assoc. Scientific Study of Consciousness, 2–6 July 2012, Brighton, UK*.
148. Dennett DC. 1988 *The intentional stance*. Cambridge, MA: Bradford Books/MIT.
149. Seeley TD *et al.* 2012 Stop signals provide cross inhibition in collective decision-making by honeybee swarms. *Science* **335**, 108–111. (doi:10.1126/science.1210361)
150. Merleau-Ponty M, Edie JM. 1964 *The primacy of perception: and other essays on phenomenological psychology, the philosophy of art, history and politics*. Evanston, IL: Northwestern University Press.
151. Gallagher S. 2005 *How the body shapes the mind*. New York, NY: Oxford University Press.
152. Hesslow G, Ivarsson M. 1996 Inhibition of the inferior olive during conditioned response in the decerebrate ferret. *Exp. Brain Res.* **110**, 36–46. (doi:10.1007/BF00241372)
153. De Zeeuw C *et al.* 1989 Ultrastructural study of the GABAergic, cerebellar, and mesodiencephalic innervation of the cat medial accessory olive: anterograde tracing combined with immunocytochemistry. *J. Comp. Neurol.* **284**, 12–35. (doi:10.1002/cne.902840103)
154. Redgrave P, Rodriguez M, Smith Y. 2010 Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nat. Rev. Neurosci.* **11**, 760–772. (doi:10.1038/nrn2915)
155. Likhtik E, Paz R. 2015 Amygdala–prefrontal interactions in (mal) adaptive learning. *Trends Neurosci.* **38**, 158–166. (doi:10.1016/j.tins.2014.12.007)
156. Nili U *et al.* 2010 Fear thou not: activity of frontal and temporal circuits in moments of real-life courage. *Neuron* **66**, 949–962. (doi:10.1016/j.neuron.2010.06.009)
157. Lisman JE, Grace AA. 2005 The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron* **46**, 703–713. (doi:10.1016/j.neuron.2005.05.002)
158. Blake D *et al.* 2006 Experience-dependent adult cortical plasticity requires cognitive association between sensation and reward. *Neuron* **52**, 371–381. (doi:10.1016/j.neuron.2006.08.009)
159. Beier KT *et al.* 2015 Circuit architecture of VTA dopamine neurons revealed by systematic input-output mapping. *Cell* **162**, 622–634. (doi:10.1016/j.cell.2015.07.015)
160. Ruff CC, Fehr E. 2014 The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **15**, 549–562. (doi:10.1038/nrn3776)

161. Apps MA, Lesage E, Ramnani N. 2015 Vicarious reinforcement learning signals when instructing others. *J. Neurosci.* **35**, 2904–2913. (doi:10.1523/JNEUROSCI.3669-14.2015)
162. Baumgartner T *et al.* 2009 The neural circuitry of a broken promise. *Neuron* **64**, 756–770. (doi:10.1016/j.neuron.2009.11.017)
163. Takahashi H *et al.* 2009 When your gain is my pain and your pain is my gain: neural correlates of envy and schadenfreude. *Science* **323**, 937–939. (doi:10.1126/science.1165604)
164. Strobel A *et al.* 2011 Beyond revenge: neural and genetic bases of altruistic punishment. *Neuroimage* **54**, 671–680. (doi:10.1016/j.neuroimage.2010.07.051)
165. Rizzolatti G, Sinigaglia C. 2010 The functional role of the parietal-frontal mirror circuit: interpretations and misinterpretations. *Nat. Rev. Neurosci.* **11**, 264–274. (doi:10.1038/nrn2805)
166. Hilgard ER. 1980 The trilogy of mind: cognition, affection, and conation. *J. Hist. Behav. Sci.* **16**, 107–117. (doi:10.1002/1520-6696(198004)16:2<107::AID-JHBS2300160202>3.0.CO;2-Y)
167. Heider F. 1944 Social perception and phenomenal causality. *Psychol. Rev.* **51**, 358–374. (doi:10.1037/h0055425)
168. Scholl BJ. 2001 Objects and attention: the state of the art. *Cognition* **80**, 1–46. (doi:10.1016/S0010-0277(00)00152-9)
169. Kovács ÁM, Téglás E, Endress AD. 2010 The social sense: susceptibility to others' beliefs in human infants and adults. *Science* **330**, 1830–1834. (doi:10.1126/science.1190792)
170. Banerjee K, Bloom P. 2014 'Everything happens for a reason': children's beliefs about purpose in life events. *Child Dev.* **86**, 503–518. (doi:10.1111/cdev.12312)
171. Kouider S *et al.* 2013 A neural marker of perceptual consciousness in infants. *Science* **340**, 376–380. (doi:10.1126/science.1232509)
172. Motzkin JC *et al.* 2015 Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. *Biol. Psychiatry* **77**, 276–284. (doi:10.1016/j.biopsych.2014.02.014)
173. Le Groux S, Verschure PFMJ. 2010 Emotional responses to the perceptual dimensions of timbre: a pilot study using physically inspired sound synthesis. In *7th Int. Symp. on Computer Music Modeling and Retrieval, Malaga, Spain.* Berlin, Germany: Springer.
174. Heywood BT, Bering JM. 2014 'Meant to be': how religious beliefs and cultural religiosity affect the implicit bias to think teleologically. *Religion Brain Behav.* **4**, 183–201. (doi:10.1080/2153599X.2013.782888)
175. Bloom P. 2012 Religion, morality, evolution. *Annu. Rev. Psychol.* **63**, 179–199. (doi:10.1146/annurev-psych-120710-100334)
176. Valdesolo P, Graham J. 2013 Awe, uncertainty, and agency detection. *Psychol. Sci.* **25**, 170–178. (doi:10.1177/0956797613501884).
177. Lorenz KZ. 1950 The comparative method in studying innate behavior patterns. *Symp. Soc. Exp. Biol.* **4**, 221–268.
178. Mathews Z, Verschure PFMJ. 2011 PASAR-DAC7: an integrated model of prediction, anticipation, sensation, attention and response for artificial sensorimotor systems. *Inf. Sci.* **186**, 1–19. (doi:10.1016/j.ins.2011.09.042)
179. Mathews Z, Cetnarski R, Verschure PFMJ. 2015 Visual anticipation biases conscious perception but not bottom-up visual processing. *Front. Psychol.* **5**, 1443. (doi:10.3389/fpsyg.2014.01443)
180. Malekshahi R *et al.* 2016 Differential neural mechanisms for early and late prediction error detection. *Sci. Rep.* **6**, 24350. (doi:10.1038/srep24350)
181. Arsiwalla XD *et al.* 2015 Network dynamics with BrainX3: a large-scale simulation of the human brain network with real-time interaction. *Front. Neuroinform.* **9**, 00002. (doi:10.3389/fninf.2015.00002)
182. Arsiwalla XD *et al.* 2015 Connectomics to semantomics: addressing the brain's big data challenge. *Proc. Comput. Sci.* **53**, 48–55. (doi:10.1016/j.procs.2015.07.278)
183. Utevsky AV, Smith DV, Huettel SA. 2014 Precuneus is a functional core of the default-mode network. *J. Neurosci.* **34**, 932–940. (doi:10.1523/JNEUROSCI.4227-13.2014)
184. Barbas H. 2015 General cortical and special prefrontal connections: principles from structure to function. *Annu. Rev. Neurosci.* **38**, 269–289. (doi:10.1146/annurev-neuro-071714-033936)
185. Wimmer RD *et al.* 2015 Thalamic control of sensory selection in divided attention. *Nature* **526**, 705–709. (doi:10.1038/nature15398)
186. Myin E, O'Regan JK. 2002 Perceptual consciousness, access to modality and skill theories: a way to naturalise phenomenology? *J. Conscious. Stud.* **9**, 27–45.
187. Simons DJ, Rensink RA. 2005 Change blindness: past, present, and future. *Trends Cognit. Sci.* **9**, 16–20. (doi:10.1016/j.tics.2004.11.006)
188. McKiernan KA *et al.* 2006 Interrupting the 'stream of consciousness': an fMRI investigation. *Neuroimage* **29**, 1185–1191. (doi:10.1016/j.neuroimage.2005.09.030)
189. Petrides M. 2007 The orbitofrontal cortex: novelty, deviation from expectation, and memory. *Ann. NY Acad. Sci.* **1121**, 33–53. (doi:10.1196/annals.1401.035)
190. Ogawa SK *et al.* 2014 Organization of monosynaptic inputs to the serotonin and dopamine neuromodulatory systems. *Cell Rep.* **8**, 1105–1118. (doi:10.1016/j.celrep.2014.06.042)
191. Sherman S, Guillery R. 2006 *Exploring the thalamus and its role in cortical function*. Cambridge, MA: MIT Press.
192. Zhang S *et al.* 2014 Long-range and local circuits for top-down modulation of visual cortex processing. *Science* **345**, 660–665. (doi:10.1126/science.1254126)
193. Liu J *et al.* 2015 Frequency-selective control of cortical and subcortical networks by central thalamus. *eLife* **4**, e09215.
194. Vico G. 1730 *Scienza nuova seconda, the new science of Giambattista Vico, revised translation of the third edition by Thomas Goddard Bergin and Max Harold Fisch*. Ithaca, NY: Cornell University Press, 1948; Cornell Paperbacks, 1976.
195. Squire LR, Dede AJ. 2015 Conscious and unconscious memory systems. *Cold Spring Harb. Perspect. Biol.* **7**, a021667. (doi:10.1101/cshperspect.a021667)
196. Henke K. 2010 A model for memory systems based on processing modes rather than consciousness. *Nat. Rev. Neurosci.* **11**, 523–532. (doi:10.1038/nrn2850)
197. Ingvar DH. 1984 Memory of the future: an essay on the temporal organization of conscious awareness. *Hum. Neurobiol.* **4**, 127–136.
198. Schacter DL, Addis DR, Buckner RL. 2007 Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.* **8**, 657–661. (doi:10.1038/nrn2213)
199. Sanders H *et al.* 2015 Grid cells and place cells: an integrated view of their navigational and memory function. *Trends Neurosci.* **38**, 763–775. (doi:10.1016/j.tins.2015.10.004)
200. Johnson A, Redish AD. 2007 Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12 176–12 189. (doi:10.1523/JNEUROSCI.3761-07.2007)
201. Buzsáki G. 2015 Hippocampal sharp wave-ripple: a cognitive biomarker for episodic memory and planning. *Hippocampus* **25**, 1073–1188. (doi:10.1002/hipo.22488)
202. Kaplan R *et al.* 2016 Hippocampal sharp-wave ripples influence selective activation of the default mode network. *Curr. Biol.* **26**, 686–691. (doi:10.1016/j.cub.2016.01.017)
203. van Fraassen B. 1980 *The scientific image*. Oxford, UK: Oxford University Press.
204. Mead GH. 1934 *Mind, self, & society*. Chicago, IL: University of Chicago Press.
205. Humphrey N. 2006 *Seeing red: a study in consciousness*. Cambridge, MA: Harvard University Press.
206. Baumeister RF, Masicampo E. 2010 Conscious thought is for facilitating social and cultural interactions: how mental simulations serve the animal–culture interface. *Psychol. Rev.* **117**, 945. (doi:10.1037/a0019393)
207. Prinz J. 2007 *The emotional construction of morals*. Oxford, UK: Oxford University Press.
208. Nisbett RE, Wilson TD. 1977 Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* **84**, 231–259. (doi:10.1037/0033-295X.84.3.231)
209. Smith JM, Szathmary E. 1997 *The major transitions in evolution*. Oxford, UK: Oxford University Press.
210. De Waal F. 2007 *Chimpanzee politics: power and sex among apes*. Baltimore, MD: JHU Press.
211. Jablonka E, Lamb MJ. 2006 The evolution of information in the major transitions. *J. Theor. Biol.* **239**, 236–246. (doi:10.1016/j.jtbi.2005.08.038)
212. McShea DW, Simpson C. 2011 The miscellaneous transitions in evolution. In *The major transitions in*

evolution revisited (eds B Calcott, K Sterelny), pp. 19–34. Cambridge, MA: MIT Press.

213. Verschure PFMJ. 1991 Chaos based learning. *Complex Syst.* **5**, 359–370.
214. Platz T *et al.* 2000 Multimodal EEG analysis in man suggests impairment-specific changes in movement-related electric brain activity after stroke. *Brain* **123**, 2475–2490. (doi:10.1093/brain/123.12.2475)
215. Cole MW *et al.* 2013 Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat. Neurosci.* **16**, 1348–1355. (doi:10.1038/nn.3470)
216. Linehan M. 1993 *Cognitive-behavioral treatment of borderline personality disorder*. New York, NY: Guilford Press.
217. Wischniewski J, Brüne M. 2013 How do people with borderline personality disorder respond to norm violations? Impact of personality factors on economic decision-making. *J. Personal. Disord.* **27**, 531. (doi:10.1521/pedi_2012_26_036)
218. Ferrarelli F, Tononi G. 2011 The thalamic reticular nucleus and schizophrenia. *Schizophr. Bull.* **37**, 306–315. (doi:10.1093/schbul/sbq142)
219. Goodkind M *et al.* 2015 Identification of a common neurobiological substrate for mental illness. *J. Am. Med. Assoc. Psychiatry* **72**, 305–315. (doi:10.1001/jamapsychiatry.2014.2206)
220. McCormick D, Thompson R. 1984 Cerebellum: essential involvement in the classically conditioned eyelid response. *Science* **223**, 296–299. (doi:10.1126/science.6701513)
221. Haber MA. 2000 Convergent inputs from thalamic motor nuclei and frontal cortical areas to the dorsal striatum in the primate. *J. Neurosci.* **20**, 3798–3813.
222. Best MD *et al.* 2016 Spatio-temporal patterning in primary motor cortex at movement onset. *Cereb. Cortex*, pbhv327. (doi:10.1093/cercor/bhv327)
223. Churchland MM *et al.* 2010 Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378. (doi:10.1038/nn.2501)
224. Fedorenko E, Duncan J, Kanwisher N. 2013 Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl Acad. Sci. USA* **110**, 16 616–16 621. (doi:10.1073/pnas.1315235110)
225. Luvizotto A, Rennó-Costa C, Verschure PFMJ. 2012 A wavelet based neural model to optimize and read out a temporal population code. *Front. Comput. Neurosci.* **6**, 1–14. (doi:10.3389/fncom.2012.00021)
226. Lallee S *et al.* 2015 Towards the synthetic self: making others perceive me as an other. *Paladyn J. Behav. Robot.* **6**. (doi:10.1515/pjbr-2015-0010)
227. Reggia JA. 2013 The rise of machine consciousness: studying consciousness with computational models. *Neural Netw.* **44**, 112–131. (doi:10.1016/j.neunet.2013.03.011)