Cogn Comput (2009) 1:50–63
DOI 10.1007/s12559-009-9007-x

Explanatory Correlates of Consciousness: Theoretical and Computational Challenges

Anil Seth

Published online: 25 January 2009
© Springer Science+Business Media, LLC 2009

Abstract Consciousness is a key feature of mammalian cognition and revealing its underlying mechanisms is one of the most important scientific challenges for the 21st century. In this article I review how computational and theoretical approaches can facilitate a transition from correlation to explanation in consciousness science. I describe progress towards identifying 'explanatory correlates' underlying four fundamental properties characterizing most if not all conscious experiences: (i) the co-existence of segregation and integration in conscious scenes, (ii) the emergence of a subjective first-person perspective, (iii) the presence of affective conscious contents, either transiently (emotion) or as a background (mood) and (iv) experiences of intention and agency that are characteristic of voluntary action. I also discuss how synthetic approaches can shed additional light on possible functions of consciousness, the role of embodiment in consciousness, and the plausibility of constructing a conscious artefact.

Keywords Consciousness · Explanatory correlate · Causal density · Complexity · Perspectivalness · Emotion · Volition · Computational model · Selfhood · Emergence

Invited article for inaugural issue of Cognitive Computation.

A. Seth (✉)
Department of Informatics, University of Sussex,
Brighton BN1 9QJ, UK
e-mail: a.k.seth@sussex.ac.uk
URL: www.anilseth.com

Introduction

Over the past 20 years experimental work addressing consciousness has shaped a new empirical science of consciousness integrating findings from psychology, neuroscience, psychiatry, neurology and cognitive science (e.g. [3, 27, 95]). Each new experimental result both enriches and constrains possible theories of consciousness and motivates further studies. At the same time, the history of science makes clear that progress is best achieved when experimental programmes are accompanied by synthetic methods which exemplify Braitenberg's law of 'uphill analysis versus downhill synthesis', the idea that complex phenomena that resist direct analysis can be better understood by analysis of less complex alternatives instantiated in simulation [9]. 'Cognitive computation' describes a class of synthetic methods highly suited for advancing the science of consciousness. The remit of cognitive computation is broad, covering biologically inspired computational accounts and models of all aspects of natural and artificial cognitive systems (Hussain, this volume). Importantly, adopting such an approach does *not* necessitate the assumption that cognitive/conscious systems are themselves computational systems; indeed, in this article no such assumption is made.

I will review several aspects of the current and future science of consciousness from the perspective of cognitive computation. These aspects are united by the development of 'explanatory correlates of consciousness': neural processes that not only *correlate with*, but also *account for* fundamental properties of conscious experience. I identify four such properties for which synthetic approaches hold particular promise: dynamical complexity, perspectivalness, emotion and mood, and volition. These properties are fundamental inasmuch as they are common to most if not

all conscious experiences. The analysis of such properties constitutes a very different approach from isolating the neural correlates of canonical experiences, such as the experience comprising only the content of 'pure red' [18]. I will conclude by discussing some wider issues raised by synthetic approaches to consciousness science. These include identifying possible functions for consciousness, assessing the role of embodiment and environmental interaction in the generation of conscious experience, and the plausibility of constructing a conscious artefact.

Explanatory Correlates of Consciousness

Basic Definitions

Consciousness is that which is lost when we fall into a dreamless sleep and returns when we wake up again. It is not a unitary phenomenon [108]. One can distinguish between conscious level, which is a position on a scale from brain-death to alert wakefulness, and conscious content, which refers to the composition of a given conscious scene at any non-zero conscious level. Conscious contents typically consist of phenomenal aspects (qualia) such as perceptual experiences (e.g. redness), bodily sensations (e.g. itchiness), emotional reactions (e.g. regret) and moods (e.g. boredom) [41]. Other conscious contents include thoughts, inner speech and usually a sense of agency, self and a subjective first-person perspective (1PP) on the world (the 'I'). Conscious level and content are related inasmuch as the range of possible contents increases with increasing conscious level.

One can also distinguish primary (sensory) consciousness from higher-order (meta) consciousness [26]. Primary consciousness reflects the presence of a 'world', of a multimodal scene composed of sensory and motor events; there is something it is like to be a primary conscious organism [64]. Higher-order consciousness involves the referral of primary consciousness to interpretative processes including a sense of self and, in more advanced forms, the ability to explicitly construct past and future scenes.

Explanatory Correlates

Conventional approaches within consciousness science have emphasized the search for the so-called 'neural correlates of consciousness' (NCCs): activity within brain regions or groups of neurons having privileged status in the generation of conscious experience [73, 95]. The ultimate aim of this approach is to discover the 'minimal neuronal mechanisms jointly sufficient for any one specific conscious percept' [50]. However, correlations by themselves cannot supply explanations, they can only constrain them.

The transition from correlation to explanation requires an understanding of *why* particular NCCs have a privileged relationship with consciousness [27, 95]. This in turn requires an understanding of key properties of consciousness that require explanation, especially those properties that are common to most or all conscious experiences. Such properties can be called structural properties [15], and the neural processes that account for these properties can be called 'explanatory correlates of consciousness' (ECCs) [81].

What are the structural properties of consciousness? A full discussion is beyond the present scope (see instead [62, 81, 82]); here I focus on four selected properties of particular relevance to cognitive computation approaches:

- Every conscious scene is both integrated (i.e. it is experienced 'all of a piece') and differentiated (i.e. it is composed of many different parts and is therefore one among a vast repertoire of possible experiences). This general property can be called 'complexity' [94]. Conscious scenes are also metastable in the sense that any given unified conscious scene shades naturally into a successive scene over a relatively stable timescale ($\sim$100 ms).
- *Perspectivalness*: The reference of conscious contents to a subjective 1PP; the existence of a 'point of view' [62]. More specifically, conscious scenes have an allocentric character, yet are shaped by egocentric frameworks. The presence of a 1PP on the world is a key component of most concepts of selfhood.
- Conscious scenes incorporate and are shaped by emotional and mood states which involve awareness of bodily processes [19, 23].
- Consciousness is marked by experiences of intention, agency, and an association with apparently voluntary action [39].

This non-exhaustive list of structural properties describes aspects or dimensions of the way the world is presented to us through conscious experience, rather than particular conscious contents per se. The current challenge for theoretical and computational models is to account for such structural properties in terms of neural system dynamics. A future goal might be to show how such properties are interdependent in the sense that accounting for one might naturally, without further assumptions, account for one or more of the others [15]. Eventually, cognitive computation models might attempt to instantiate these properties in the service of creating a conscious artefact.

Consciousness, Complexity and Causal Density

Consciousness and Complexity

The association of consciousness with complexity in the form of the coexistence of integration and differentiation represents a fundamental insight into conscious experience. It is at the heart of two related theories of consciousness, the 'dynamic core hypothesis' [27] and the 'information integration theory of consciousness' [93], both of which emphasize that the complex nature of consciousness is highly informative for the organism, in the specific sense that the occurrence of any particular conscious scene rules out the occurrence of a very large repertoire of alternative possibilities [94]. The structural association of consciousness with complexity provides an attractive opportunity for developing a corresponding explanatory correlate. Such a correlate would consist in a description of neural dynamics exhibiting high simultaneous integration and differentiation.

Measures of Dynamical Complexity

Several candidate descriptions have been proposed that characterize quantitatively the co-existence of integration and differentiation in multiple simultaneously recorded time series. These include 'neural complexity' [97], 'information integration' [93] and 'causal density' [76, 77]. Detailed theoretical comparisons of these measures can be found in [82, 83]; here I describe only their basic properties and differences (Fig. 1).

- Neural complexity expresses the extent to which a system is both dynamically segregated, so that small subsets of the system tend to behave independently, and dynamically integrated, so that large subsets tend to behave coherently. Formally it is equal to the sum of the average mutual information across all bipartitions of a system [97], where mutual information measures the uncertainty (entropy) about one system (or subset) that is accounted for by observations of another.
- Information integration (Φ) has been proposed as a way to quantify the total amount of information that a conscious system can integrate [93]. It is defined as the 'effective information' across the informational 'weakest link' of a system, the so-called 'minimum information bipartition'. Effective information is calculated as the mutual information across a partition in the case where outputs from one subset have maximum entropy, and the minimum information bipartition is that partition of the system for which the effective information is lowest.
- Causal density is a global measure of causal interactivity that captures dynamical heterogeneity among elements (differentiation) as well as their global dynamical integration [76, 77]. It is calculated as the fraction of interactions among elements that are causally significant, according to a statistical interpretation of causality introduced by Granger [36]. According to 'Granger causality', a variable A 'causes' a variable B if past observations of A help predict B with greater accuracy than possible by past observations of B alone. Granger

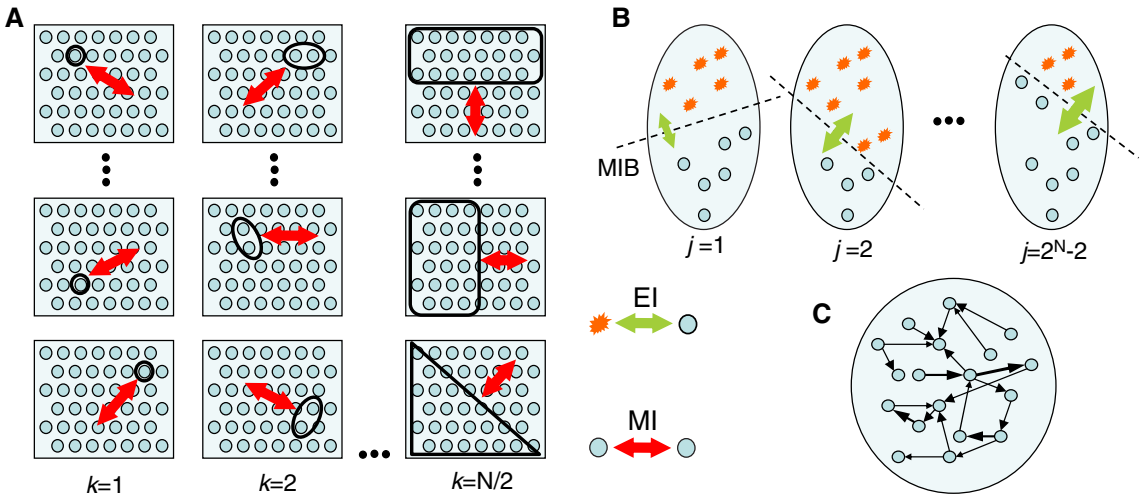

Fig. 1 Measuring complexity for a neural system X composed of N elements. **a**. Neural complexity (C_N) is calculated as the sum of the average mutual information (MI) over $N/2$ sets of bipartitions indexed by k (e.g. for $k = 1$ an average MI is calculated over N bipartitions). **b.** Information integration (Φ) is calculated as the effective information across the 'minimum information bipartition' (MIB). To calculate effective information for a given bipartition (indexed by j), one subset is injected with maximally entropic activity (*stars*) and MI across the partition is measured. **c**. Causal density is calculated as the fraction of interactions that are causally significant according to Granger causality. A weighted (and unbounded) version of causal density can be calculated as the summed magnitudes of all significant causal interactions (depicted by *arrow width*). Reprinted with permission from [83] (Refer online version for colour figure)

causality is easily extensible to multivariate situations and is usually implemented through linear autoregressive modelling, though non-linear extensions exist. High causal density indicates that elements within a system are both globally coordinated in their activity (in order to be useful for predicting each other's activity) and at the same time dynamically distinct (so that different elements contribute in different ways to these predictions).

Having explicit measures of dynamic complexity can transform a *property* of consciousness (integrated and differentiated experience) into a *criterion* that can be applied to empirical or simulation data. This transformation is at the heart of the strategy of developing ECCs [15]. In addition, different measures can operationalize subtly different aspects of the same overarching property. For example, unlike neural complexity and causal density, Φ is explicitly cast as a measure of the capacity of a system to generate complex dynamics, as opposed to a measure of dynamics per se. This is a critical difference in view of the corresponding 'information integration theory of consciousness' [93] which proposes that consciousness is itself a capacity rather than a process.

New measures can also correct perceived deficiencies in previously proposed measures. For example, unlike neural complexity, both causal density and Φ are sensitive to causal interactions among elements of a system (mutual information is a symmetric measure, whereas Granger causality and effective information are directed). This is important inasmuch as neural dynamics implement causal interactions. Causal density is also sensitive to dynamics that are smeared out over time, depending on the number of 'lags' incorporated into the underlying autoregressive model (Fig. 2). In contrast, both neural complexity and Φ are based on representations of dynamics derived through zero-lag correlations; these measures are therefore insensitive to temporally smeared dynamics. Finally, both causal density and neural complexity are calculable in practice for non-trivial systems, whereas Φ can at present only be calculated for simple models exclusively in simulation [82, 96].

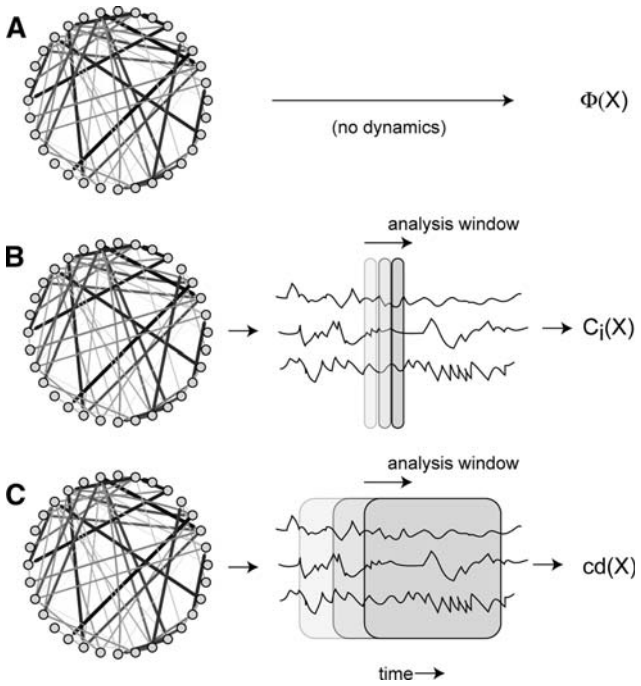

Fig. 2 Incorporation of time by measures of dynamical complexity for a neural system *X*. **a** Information integration is a static measure; it can be measured by assuming Gaussian dynamics which allows analytic calculation directly from the network anatomy [96]. **b** Neural complexity can be either a static measure or a dynamic measure. Analytic calculation of neural complexity can be accomplished in the same way as for information integration, yielding a measure of capacity. Alternatively, neural complexity can be calculated on the basis of the recorded activity of a network, yielding a measure of 'interactive complexity'. However, interactive complexity is sensitive only to 'thin' dynamics, since it is calculated on the basis of zero-lag temporal correlations. **c** Causal density is by definition a dynamic measure, since it reflects Granger causality interactions that depend on multivariate modelling of network dynamics. Causal density is sensitive to temporally smeared dynamics because a given multivariate model will reflect temporal interactions across a time period that depends on the number of 'lags' incorporated into the model

Simulation Models

Differences among measures can be tested using simulation models. Recently, Shanahan [87] compared neural complexity and causal density in a computational model of spiking neurons arranged in loosely connected clusters. As clusters became more strongly interconnected, causal density showed a peak at an intermediate point characterized by sustained, desynchronized yet partly integrated spiking activity. In contrast, neural complexity only began to detect complexity when activity in clusters began to synchronize, at exactly the point where the *dynamical* complexity of the network started to diminish according to causal density. As suggested above, the likely explanation for this is that neural complexity is insensitive to integration or segregation that is smeared over time.

A related modelling approach involves developing model neural systems that are tuned to show high values of a given measure and then comparing their structure with aspects of neuroanatomy thought to underlie consciousness. For example, neural complexity has been shown to be high for networks that show structural similarity to mammalian thalamocortical networks, in that both have small-world network characteristics [89]. Small-world networks, which consist of loosely coupled sets of highly interconnected clusters, show many interesting dynamical properties including high synchronizability, enhanced signal propagation speed, low overall 'wiring length' and high

robustness to damage [102]. The Shanahan study described above explored one particular method for implementing small-world networks.

Synthetic models can also explore the functional utility of high dynamical complexity in neural systems. For example, both neural complexity and causal density have been shown to increase as the behavioural flexibility of a simple artificial agent increases [76, 80, 106]. Sporns and Lungarella [88] showed that neural networks optimized for high neural complexity behaved successfully in a target-reaching task, despite the fact that target-reaching behaviour had not been explicitly selected for. These findings are consistent with the idea that dynamical complexity (and therefore perhaps consciousness) can provide adaptive advantages during behaviour in virtue of facilitating response flexibility.

Future Challenges

Looking ahead, there is a need for new measures and models that capture 'metastability' in neural systems, which refers to simultaneous integration and differentiation in the time domain [10, 104]. Metastability is a deep structural property of consciousness in that each conscious moment is constituted by a rich interweaving of the present, the immediate past (retension) and the predicted future (protension) [26, 48, 92]. Such models and measures might be most likely to arise through the tools of dynamical systems theory which give special attention to non-linear and transitory aspects of system dynamics [16]. In addition, synthetic models can explore relations between complexity-based accounts and approaches which tend to emphasize the integrative nature of consciousness rather more than its differentiated aspects. These include global workspace theory [2, 24] and the notion that consciousness is mediated by synchronized neural activity [30] or by local and global patterns of reentry [53]. Synthetic approaches can also analyse the dynamical complexity properties of models incorporating detailed neuroanatomy thought to underlie consciousness in mammals. Suitable points of departure include large-scale models of thalamocortical networks [47] and models of the 'structural core' within mammalian cortex which comprises hubs of particularly dense interconnectivity among certain medial, parietal, temporal and frontal cortical modules [40].

Perspectivalness

First-Person Perspectives

Our conscious mental life generally has a point-of-view, a subjective phenomenal 1PP located somewhere between and behind the eyes and imparting an egocentric component to conscious contents. 1PPs are not always like this; they can shift spatial location in autoscopic and out-of-body experiences and they may be absent entirely in deep meditative states.[1] A 1PP is an essential part of what in folk psychological terms is a 'self'. However, although there may be no such things as selves in the world, the *experience* of being a self does exist [62]. The normal presence of a 1PP—'perspectivalness'—is therefore a structural property of consciousness requiring explanation. It is worth distinguishing basic perspectivalness from the ability of some organisms (notably humans) to understand the world from the point-of-view of another. This competence—sometimes referred to as 'theory of mind'—may require perspectivalness, but the converse is unlikely to be true.

Thomas Metzinger's influential 'self-model theory of subjectivity' proposes that a 1PP originates through the operation of a self-model, an "episodically active representational entity whose content is determined by the system's very own properties" [63, p. 218], the purpose of which is to regulate the system's interactions with its environment. The existence and causal efficacy of human self-models in some form has been extensively demonstrated empirically. For example, the experience of a phantom limb that can follow amputation, and the alleviation of the 'phantom pain' in this phantom limb by providing false cross-modal feedback [72] indicate the existence of a self-model and show its potential for recalibration. Even more dramatic is somatoparaphrenia, a syndrome characterized by delusions of disownership of left-sided body parts [99]. Disturbances of 1PPs themselves are also evident in out-of-body and autoscopic experiences. Interestingly, such experiences can be induced experimentally by a combination of virtual reality and multimodal feedback [29, 55] again showing the rapid adaptivity of aspects of biological self-models.

Simulation Models

The notion that 1PPs originate in the operation of self-models invites synthetic modelling. Synthetic self-models can be implicit in predictions of sensorimotor flow, or they can be explicit. In the former category, Grush [38] has described a framework based on forward modelling and Kalman-filter signal processing in which neural circuits act as models of body–environment interactions. These models are driven by efference copies of motor commands and provide expectations of sensory feedback, and they can be run off-line in order to produce imagery and evaluate the

[1] Autoscopy is the experience of seeing one's own body in extrapersonal space, whereas an out-of-body experience is characterized by a shift in perspective to a location outside the body [55].

outcomes of different actions. Similarly, Revonsuo [74] argues that consciousness involves a neural 'virtual reality' apparatus allowing off-line simulations of potential threats, and Hesslow [42, 43] has also proposed a model of consciousness in terms of simulation. A minimal robotic implementation of implicit self-modelling has been described by [110].

Perhaps the most explicit example of the development of an artificial self-model is provided by Bongard et al. [7], who use artificial evolution techniques (genetic algorithms) to enable a four-legged 'starfish' robot to generate autonomously a representation of its own body (Fig. 3a). This robot is capable of re-adapting its self-model following damage (e.g. removal of a leg) and can run its model as an internal simulation in order to discover control strategies leading to effective locomotion. This example shows that a self-model need not be conscious, allowing that even in humans unconscious self-models may shape perspectivalness in conscious contents. It also shows that self-modelling, like consciousness itself, is a process and not a 'thing' [48]. Another example of explicit self-modelling is Holland's 'Cronos' robot, which consists of a complex anatomically detailed humanoid torso and a correspondingly complex simulation model of self and world (Fig. 3b; [45]). Inspired by Metzinger's theory, this study explores the implications of the view that animals regarded as intelligent (and perhaps conscious) tend to have complex body morphologies and interact with their environment in correspondingly rich ways.

Current challenges for consciousness science involve building on the above work in a number of ways. Explicit self-model development such as that described by Bongard et al. [7] needs to be cashed out in terms of neural mechanisms rather than at the level of genetic algorithms and physics-engines. This would generate predictions about the underlying neurophysiology and could connect with the extensive body of work dealing with forward and inverse modelling for motor control in neuroscience [60, 105]. More fundamentally, the mechanisms by which a 1PP emerges from self-modelling need further elaboration. This will likely involve explaining the interaction between the egocentric framework of a 1PP and the allocentric character of the perceived world. Although we perceive the world from a particular point of view, the experienced world consists of objects in relation to each other; in other words our everyday conscious experience seems to be intermediate between allocentric and egocentric frameworks [57]. Such models may draw on experimental work suggesting distinct neuroanatomical loci for egocentric and allocentric maps and their interactions [11]. Synthetic self-models also need to account for the various manipulations and disturbances of perspective and self-representation that can be induced experimentally or that occur following brain damage, surgery or psychiatric disturbances. Finally, there are opportunities to explore how off-line operation of self-models can be used to guide behaviour by simulating the consequences of different actions, providing insight into the unconscious operation of self-models and into conscious off-line modes such as dream states and imagery.

Future Challenges

In the long-term, synthetic models could elaborate the hypothesis that self-models arose in evolution through the need to control increasingly complex body morphologies having progressively more degrees of freedom [20]. This hypothesis is of particular relevance when considering the possibility of consciousness in non-mammalian creatures with large brains and rich morphology such as the octopus [28]. The relation between self-modelling and perspectivalness could also be expanded to include other aspects of the concept of the self. In addition to perspectivalness, Metzinger identifies two further target properties: 'mineness', a non-conceptual sense of ownership of experiences, thoughts, feelings, body parts, etc., and 'selfhood' or 'continuity', the experience of being a self and of being

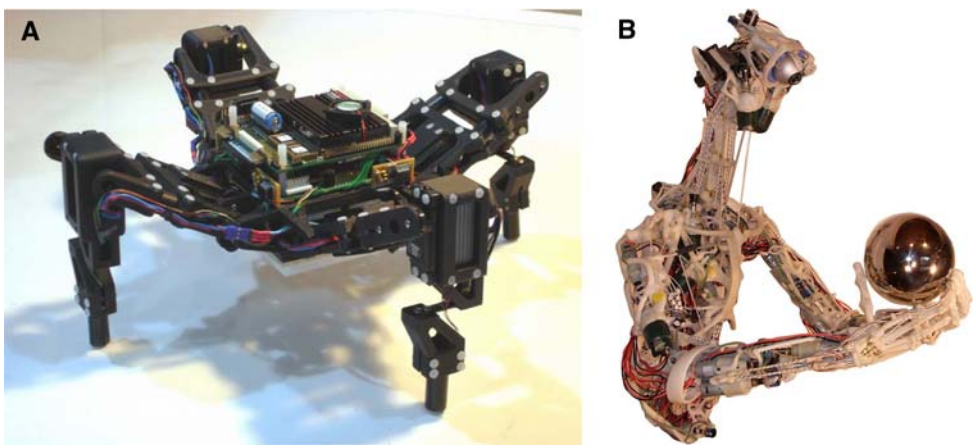

Fig. 3 **a** Starfish, a physical robot with eight motorized joints, eight angle sensors and two tilt sensors. **b** Cronos, an anthropomimetic robot inspired by human skeletal structure and musculature. Permissions from (**a**) Cornell University and Josh Bongard, Victor Zykov, and Hod Lipson, and (**b**) Owen Holland and The Robot Studio

more-or-less identical across time [63]. These are both structural aspects of normal conscious experience which may derive from interactions between self-modelling and other neural systems including those responsible for affective processing. Indeed, such interactions are prominent in the study of Damasio [23], who hypothesizes the existence of an 'as-if body loop' as a variety of self-modelling in which the brain internally simulates emotional body states, as discussed below.

Emotional Consciousness

Emotion, Cognition and Consciousness

Cognition and emotion are tightly coupled in the brain and in behaviour, and their interactions are now beginning to be unravelled in detail [68]. Performance on standard cognitive tasks, especially those involving decision making, can be severely impaired after the loss of emotional responses [22] and brain regions viewed as 'cognitive' and 'emotional' are highly integrated within the brain [68]. The relation between emotion and consciousness is less well understood. Conscious experiences generally involve emotional (affective) components, both transiently (e.g. experiences of rage, of delight) and as a temporally extended background or mood (e.g. sadness). These components, or 'feelings', can interact with other conscious contents (perceptions, thoughts), and brain structures important for emotion (e.g. brainstem nuclei and certain midline cortices) overlap with structures that regulate conscious level [98]. However, it is not known whether basic emotional processing is necessary for conscious experience, and experimental methods for dissociating emotional conscious contents from inducing stimuli are poorly developed in comparison to visual neuroscience methods.

An influential idea, originated by William James, proposes that emotional experiences are mediated by interoceptive representations of changes in bodily state (as opposed to exteroceptive perceptions of external stimuli). In other words, feelings are constituted by perceptions of internal processes such as heartbeat and vasomotor activity [17, 19, 23, 67]. Extensions to this idea suggest that the experience of a 'core self' originates via interoceptive representations of the body, both in terms of its morphological properties as discussed above (see section "Perspectivalness") and in terms of its internal physiological milieu [23]. Importantly, this 'core self' is distinct from the concepts of a metacognitive, narrative or reflective self, and corresponds to the explanatory targets of 'mineness' and 'continuity' that constitute, along with perspectivalness, a basic instantiation of selfhood.

Several general theories of consciousness emphasize a Jamesian emotional component. Damasio's 'somatic marker hypothesis' proposes that core (primary) consciousness arises via non-verbal representations of how an organism's internal state is affected by the perception of an external object, where this representational process helps to place the perceived object in a salient spatiotemporal context [23]. Damasio's framework includes an 'as-if body loop' which involves simulation of interoceptive data, providing a connection to the predictive self-modelling concepts described above (see also [109]). In Edelman's 'theory of neuronal group selection', conscious experiences depend on re-entrant interactions between brain regions supporting current perceptual categorization and those responsible for a 'value-category' memory, where 'value' reflects the operation of pleasure, pain and other emotional salience networks [26]. Other theoretical treatments of emotion and consciousness are provided by Lambie and Marcel [52] who emphasize distinct modes of attention to emotional contents, and Panksepp [67] who argues that mechanisms of basic emotional consciousness are likely to be strongly conserved among all mammalian species.

Simulation Models

Synthetic models have both contributed to and exploited our increasing understanding of the links between cognition and emotion. For example, 'affective robotics' describes attempts to enhance adaptive behaviour through emotional modulation of decision making, to facilitate human–robot interactions by exploiting human receptivity to emotional stimuli, as well as to enhance our understanding of the neuromodulatory interactions underlying emotional processing per se [25, 31, 107]. Disembodied simulation models of emotion have also become increasingly prominent. Such models however have so far focused mainly on fear and reward systems, modelling neural interactions involving the amygdala and frontal cortices [37, 101]; for a more general approach see [35].

The synthetic modelling work directly addressing emotional consciousness is scarce. Thagard and Aubie [90] describe a model involving multiple interacting brain regions integrating perceptions of bodily state with cognitive appraisals of current situations; Shanahan [86] has augmented a 'global workspace' model with an affective component in order to mediate action selection, and Bosse et al. [8] formalize the aspects of Damasio's somatic marker hypothesis including the 'as-if body loop'. But no synthetic work to date describes an explanatory correlate of emotional consciousness to the extent that has been possible with respect to the complexity of experience and, to a lesser extent, the origin of a 1PP.

Future Challenges

One avenue for further progress consists in developing increasingly sophisticated models of neural systems and processes thought to be involved in emotional consciousness, shaped by the framework of interoceptive awareness. Recent studies have highlighted the importance of right insula cortex in interoceptive perception (e.g. [19]), with anterior subregions possibly involved in explicit representations of feeling states that may underlie higher-order representations of self, extending beyond the 'core'.[2] The so-called 'default network' of the brain may also be implicated in emotional and self-related processing [71]. Activity in this network is correlated with stimulus-independent thought and with interoceptive and self-related conscious content [58], and is anticorrelated with sensitivity to external somatosensory stimuli [6]. Moreover, posterior components of this network are part of the 'structural core' described earlier [40]. Further modelling exploring dynamical properties of these networks in the context of interoceptive processing and self-modelling is likely to be very valuable.

A second and more challenging approach comprises a continuing search for explanatory correlates of emotional consciousness. This search could be guided by the notion of selfhood. As mentioned above, core selfhood incorporates perspectivalness, mineness and continuity, with the latter two properties also appearing as defining properties within affectively grounded theories of the self [65]. Further development of synthetic self-models may therefore shed new light on emotional aspects of self. Empirical data may also help identify proper explanatory targets for emotional consciousness. For example, patients with depersonalization disorder (DPD) show reduced activity in insula cortex [69]. In contrast to autoscopic and out-of-body experiences, DPD does not involve changes in point-of-view, but instead involves a striking lack of subjective validity for perceptions, thoughts, memories and self-consciousness. Thus, affective components of consciousness may be those that impart validity and perceived reality to our experiences. The extent to which these components overlap with 'mineness' and 'continuity' is not yet clear.

Volition and Downward Causality

Voluntary Action and 'Free Will'

The idea that consciousness functions to initiate voluntary action is prominent in folk concepts of consciousness, even though it has been widely challenged both empirically [56] and theoretically [103]. Nonetheless, as with the self, even though 'free will' may not exist in the world, the *experience* of volition certainly does exist and therefore requires explanation.

Daniel Wegner's influential theory of 'apparent mental causation' predicts *when* experiences of volition might occur [103]. According to this theory, we experience volition when conscious mental content is inferred, rightly or wrongly, to have produced the corresponding physical action. Such inferences are made only when the following constraints are satisfied: (i) primacy (the mental content immediately precedes the action), (ii) consistency (the content corresponds to the action) and (iii) exclusivity (there is no other plausible causal factor). Although there is experimental evidence in support of this theory, no explanation is given for the qualitative character of experiences of volition; in other words the theory does not propose a corresponding explanatory correlate.

Complementing psychological theories such as Wegner's are new data shedding light on the neural mechanisms underlying voluntary action. Experiments on volition typically consider voluntary action to reflect a 'freedom from immediacy' in terms of responses to environmental stimuli [84]. In a recent review, Haggard [39] has described a network of premotor, cingulate and frontal brain regions that are distinctively implicated in voluntary action. One area in particular, the pre-supplementary motor area (preSMA) seems to be especially critical both for experiences of volition and for expression of voluntary action.[3] Haggard also offers a general model of human volition as a sequence of decision processes of increasing specificity, from early 'whether' decision that involve motivations to late predictive checks generating possible vetoes (Fig. 4). But again, correlations between activity in particular brain regions and conscious contents do not by themselves account for the qualitative nature of that content.

What might an explanatory correlate of voluntary experience look like? Experiences of volition are characterized both by *intention* (the 'urge' to perform an action) and *agency* (the feeling that the intended action has caused something in the body and/or world to take place). A naïve interpretation of these features is that conscious experiences are distinct from their physical substrates and yet cause physical events, in the brain or elsewhere. This position assumes dualism and will not be discussed further. More satisfactory is the proposal that voluntary actions lie at one end of a continuum whose other extreme is defined

[2] Intriguingly, the thalamocortical pathway conveying detailed interoceptive signals to the right anterior insula appears to be unique to primates [17].

[3] As Haggard emphasizes, activity in preSMA is not to be interpreted as the origin of 'free will' in the sense of an uncaused cause. Brain circuits underlying volition likely consist of complex loops, and indeed input to preSMA from basal ganglia is thought to play an important role in the generation of voluntary action.

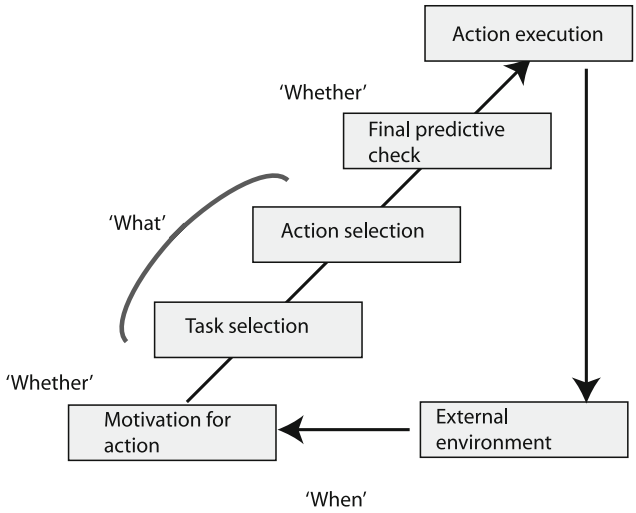

Fig. 4 Haggard's model of human volition. Volition is interpreted as a set of decisions of increasing specificity. 'Whether' decisions are made both early (motivation dependent) and late (final check), and 'what' decisions manage different levels of action specification. The timing of voluntary action ('when' decisions) depends on both environmental stimuli and internal motivational states. Adapted from [39]

by stimulus-driven simple reflexes; this is the 'freedom from immediacy' noted above [39, 84]. Consistent with this proposal, and in contrast to dualism, is the notion that consciousness is *entailed* by certain brain activity patterns, in the same way that the molecular structure of haemoglobin entails a particular spectroscopic profile [27]. On this view, certain physical events (in the brain and elsewhere) could not occur without the corresponding conscious experience even though the conscious experience itself is not changing the course of the underlying physical events. It therefore remains legitimate to speak of consciousness causing physical events (and successive conscious experiences) for the simple reason that it could not be otherwise; the physical event could not happen in the absence of the corresponding conscious experience. Putting these ideas together, we arrive at the notion that an experience of volition consists in a conscious experience with phenomenal features of intention and agency, entailed by neural activity mediating action not directly determined or very indirectly determined by external stimuli.

Simulation Models of Volition

Synthetic models of voluntary action are scarce, and models elaborating explanatory correlates of volitional experience are completely lacking. Among the former, existing models have addressed different aspects of Haggard's conceptual scheme (Fig. 4). For example, Cisek [14] has proposed that prefrontal signals reflecting task selection bias action selection processes mediated within parietal-premotor circuits. A related model suggests that frontopolar cortex enables the concurrent maintenance of two competing tasks (goals), updating the value of each task as reward expectations change [51]. A wealth of models tackle the relatively constrained problem of action selection; however, it is beyond the present scope to provide a review (see [70] for a selection). Further development of computational models of volition—independently of any association with consciousness—remains an important challenge [39].

Towards an Explanatory Correlate of Volitional Experience

A key challenge is to develop an explanatory correlate of intention and or/agency that is consistent with the functional aspects of voluntary action. I propose an approach based on the notion of 'emergence'. An emergent process or property is a macroscopic property that is somehow 'more than the sum' of its component parts. For example, a flock of starlings wheeling in the sky prior to roosting seems 'more than the sum' of the trajectories of the individual birds.[4] According to the concept of 'strong emergence', a macro-level property is *in principle* not identifiable from micro-level observations. Furthermore, strongly emergent macro-level properties are often assumed to have 'downwardly causal' influences on micro-level properties [49].

David Chalmers has made explicit the recurring idea that there is only one example of strong emergence in nature, and that is consciousness [12]. Two intuitions appear to drive this idea. First is the suspicion that even complete knowledge of the physical interactions sustained by brains will not provide an understanding of what it is like to have a conscious experience. This reflects the infamous 'hard problem' of consciousness, and it is precisely to defuse the apparent intractability of this problem that the concept of an ECC has been introduced, here and in related terminology elsewhere. Second is the notion that conscious experiences have causal efficacy in the world. This maps cleanly onto the notion of downward causality in strong emergence inasmuch as a conscious experience of volition might be a strongly emergent property having downwardly causal influences on its underlying neural activity, with subsequent causal chains spreading out to the body and the environment.

The concept of strong emergence is however problematic. The claim that the macro is in principle not identifiable from the micro rejects mechanistic

[4] Strictly speaking this is a description of 'property emergence'. There is also the notion of 'temporal emergence' which refers to the appearance of a qualitatively new phenomenon over time.

explanations altogether, apparently calling a halt to scientific advance in the absence of new fundamental principles of nature [12]. The notion of downward causality is also metaphysically awkward. It contravenes the plausible doctrine that 'the macro is the way it is in virtue of the way things are at the micro', an idea that has been expressed variously as 'causal fundamentalism' or 'supervenience' [49]. It also raises the challenge of how to resolve conflicts between competing micro- and macro-level causes [4].

A useful alternative to strong emergence is provided by the notion of 'weak emergence', which proposes that macro-level properties are derived from the interaction of micro-level components but in complicated ways such that the macro-level property has no simple micro-level explanation [4]. It is possible to operationalize weak emergence such that a macro-property is weakly emergent *to the extent that* it is difficult to identify from micro-level observations [78]. This definition requires an objective measure of the non-triviality of micro-to-macro inferential pathways, as well as a means of verifying micro-to-macro causal dependence. I have recently described such a measure, 'G-emergence', which quantifies the extent to which a macro-level property is simultaneously (i) *autonomous from* and (ii) *dependent upon* its underlying causal factors [78]. This measure is implemented using the statistical framework of non-linear Granger causality and offers a metaphysically innocent means of characterizing downward causality simply as the Granger causality from macro-variable(s) to micro-variable(s).

By considering conscious experiences as weakly emergent—rather than strongly emergent—from their underlying neural mechanisms, downward causality could provide a useful explanatory correlate of experiences of volition. Specifically, one can hypothesize that the extent to which a conscious experience includes a volitional component will correlate with measurable downward Granger-causality from macro-level descriptions of brain dynamics relevant to consciousness to micro-level brain descriptions.

A challenge for this approach is that it is not clear what would constitute a relevant macro-level variable given the impossibility of recording first-person experience except through a behavioural report. Candidates might include synchronized activity in neural implementations of a global workspace, or in the 'default network', or in activity in all or part of the structural or dynamic core. Alternatively, one might look for causal influences extending from specific neural structures implicated in volition such as the preSMA. In general, however, it should be possible to identify relevant macro-level variables directly from micro-level data. Beginning with Amari [1], various approaches under the rubric of 'statistical neurodynamics' have addressed this problem. Shalizi and Moore [85] define a macro-state as one that has higher 'predictive efficiency' than the micro-variables it derives from, in which predictive efficiency is based on Crutchfield's [21] concept of an epsilon-machine. Bishop and Atmanspacher [5] introduce the concept of 'contextual emergence', proposing that macro-level properties consist in 'stability criteria' which constrain (or 'enslave') the interaction of micro-level components; they give the example of Bénard convection currents which appear to govern the role of individual molecules in a liquid.

Despite the difficulties involved in identifying relevant macrostates, it is likely that formal frameworks describing consciousness as emergent from underlying neural dynamics will be useful as a component within synthetic and theoretical approaches. Rather than attempting to utilize such frameworks to solve the (hard) problem of consciousness *tout courte*, it may be more productive to leverage multi-level theoretical constructs such as downward causality to define explanatory correlates for specific dimensions of conscious experience, in particular the experience of volition.

Discussion

Summary of Contributions

'Cognitive computation' approaches can contribute to the science of consciousness in at least two interacting ways. First, the construction and analysis of synthetic models (software and/or robotic) can help connect neural dynamics to structural properties of conscious experience. Second, theoretical approaches can define ECCs, whose properties and experimental predictions can be explored through the subsequent construction of synthetic models. Importantly, these approaches do *not* assume that cognitive/conscious systems are themselves computational; no such assumptions have been made in this article.

I have described four challenges for synthetic and theoretical approaches to consciousness science. The first is the design of new quantitative measures reflecting the dynamical complexity of conscious experience. The importance of possessing such measures is hard to overestimate: the history of science has demonstrated repeatedly that the ability to measure a phenomenon is an important stage in the evolution of its scientific understanding [13]. Moreover, reliable measures will not only enhance basic scientific understanding but will be useful in practical contexts including the assessment of conscious level in brain-damaged patients and perhaps in non-human animals. The second challenge involves developing models of the emergence of 1PPs from internal predictive self-models. Responses to this challenge, taken together with the development of ECCs of emotional components of conscious

experience (the third challenge), promise substantial new insights into self-related components of normal human consciousness and, more prospectively, into self-related disorders such as schizophrenia and depersonalization. The final challenge is also related to the self. Experiences of volition are fundamental to selfhood, and explanatory correlates of intentionality and agency may leverage both new information about underlying neural mechanisms and new theoretical entities such as operational definitions of weak emergence and downward causality.

I will end by discussing briefly some wider issues raised by computational approaches to modelling consciousness.

Functions of Consciousness

Specifying a plausible function (or functions) for consciousness has proven remarkably difficult. The extreme positions that consciousness plays no causal role ('epiphenomenalism') or that any cognitive/behavioural activity can in principle be carried out without consciousness ('conscious inessentialism') are counterintuitive but hard to disprove [79]. 'Cognitive computation' approaches can address possible functions by implementing underlying mechanisms in concrete models. For example, existing models have implemented varieties of global workspace architectures in order to elaborate the hypothesis that consciousness serves to integrate otherwise independent cognitive and neural processes [2]. These models have replicated certain experimental phenomena such as the attentional blink [24], and have been extended to include action selection and deliberation [32] and have incorporated simple internal simulation properties [86].

Other models are starting to address possible functions of highly complex (or causally dense) dynamics. For example, networks with high neural complexity or causal density show increases in behavioural flexibility in challenging environments (see section "Simulation models"). Future models might explore the more general hypothesis that the complex neural dynamics underpinning consciousness provide adaptive discriminations, in the sense that the occurrence of any given conscious scene rules out the simultaneous occurrence of a vast number of alternative experiences [94]. Models of perspectivalness, emotion and volitional experience also shed light on possible functions of consciousness. The interplay between egocentric and allocentric representations in the emergence of a 1PP may supply a stable arena for actions [61], emotional aspects of consciousness may provide adaptive biases in decision making [22] and the mechanisms underpinning volitional experiences may implement exploratory, goal-directed and 'immediacy-free' actions [39].

The success of current and future models of conscious functions can be judged by the extent to which (i) the modelled neural processes provide useful functionality that is otherwise difficult to account for and (ii) the models generate testable experimental predictions. Importantly, most synthetic models address only so-called 'causal role' functions (i.e. what does consciousness do?) rather than phylogenetic functions (i.e. why did consciousness evolve?). Establishing phylogenetic functions is in general harder than testing for causal role functions [34], but causal role functions can at least suggest plausible hypotheses with respect to the evolution of consciousness.

Embodiment

A useful avenue for exploring causal role functionality is to build synthetic models in which the proposed ECCs are embodied in simulated or robotic bodies that interact with external environments. Opinions differ as to whether embodied and environmentally embedded sensorimotor interactions are necessary [66] or not necessary [95] for conscious experience. Dream states (and 'locked in' states [54]) show that conscious experiences are possible in the absence of body–environment interaction; however, the dreaming or locked-in brain still has a body and it is plausible that a history of brain–body–environment interactions is needed for conscious experience during waking or sleeping. In any case, normal human consciousness is implicated in guiding behaviour and its contents during waking are continually shaped by brain–body–environment interactions.

Embodied synthetic models are particularly salient with respect to perspectivalness and the emergence of basic selfhood. Although it is possible to envisage a disembodied complex system having high dynamical complexity, it is difficult to conceive that anything like a 1PP could exist in the absence of the spatiotemporal context provided by a body. Embodied synthetic models therefore provide ideal grounds for elaborating both ECCs of perspectivalness and for testing more general theories of consciousness that emphasize predictive self-modelling [42]. An implication of such theories is that perspectivalness and/or selfhood may depend on a sufficiently rich morphology supporting complex agent–environment interactions. Holland's study [45] shows such rich morphologies are now available, both in hardware using novel engineering paradigms based on human anatomy and in software using physics engines to design arbitrarily realistic body shapes and interactions (Fig. 3b).

More generally, it is possible that embodiment is significant for consciousness inasmuch as conscious organisms display a drive towards maintaining physiological integrity. In other words, organisms 'care' about the viability of their bodies and this 'caring' may be manifest in consciousness through motivations, moods and other emotional conscious content. This view points to a

continuity between 'life' and 'mind' in terms of patterns of organization, suggesting that a satisfactory theory of consciousness will need to be grounded in metabolic and physiological homeostatic and homeodynamic processes [59, 91, 100, 109]. Finally, Thompson and Varela [92] advocate an 'enactive' view according to which processes crucial for consciousness cut across the brain–body–world divisions, and are not brain-bound neural events.

Towards a Conscious Artefact

This article has described synthetic approaches for modelling key processes underlying consciousness, with the objectives of gaining insight into these processes and their consequences, and promoting conceptual clarification and development. An alternative goal is that of *instantiating* consciousness through the implementation of mechanisms underlying its key properties. The distinction between these two goals is captured by the notions of 'weak' versus 'strong' approaches to 'artificial consciousness', where the former aims at simulation and the latter at instantiation [15, 44].

The weak/strong distinction is manifest in the other sciences of the artificial, namely artificial life and its original context, artificial intelligence [75]. In both cases, proposed examples of instantiation in the form of computational models or robotic devices have remained hotly disputed. However, the possibility of instantiating full-blown intelligence or life in an artefact is not mysterious in principle. For example, although it is increasingly accepted that computational models of life are indeed models in the weak sense, there is now a new and overlapping field—synthetic biology—in which researchers create new life forms by the artificial synthesis of genetic material and the subsequent implantation of this material into surrogate embryos [33]. The consensus here is that these new organisms are in fact alive and are not merely models.

Is it possible to envisage an artefact endowed with full-blown consciousness? One possibility is that future progression in weak (simulation) artificial consciousness may inevitably lead towards a strong version (instantiation) [15]. As one successively builds in new constraints to match objections that become apparent through the building of models, so the models in question may actually tend towards the instantiation of systems that might genuinely be considered conscious. It is not yet clear whether a model of consciousness sufficiently rich to account for all its structural properties will turn out to be implementable in computers or robots. In line with synthetic biology it might instead be that such 'models' will require implementation in neural or some other materials.

Acknowledgements Preparation of this article was supported by EPSRC leadership fellowship EP/G007543/1. I am grateful to Tom Ziemke for useful comments on a first draft and to Owen Holland for Fig. 3b.

References

1. Amari S-I. A method of statistical neurodynamics. Kybernetik. 1974;14:201–15.
2. Baars BJ. A cognitive theory of consciousness. New York: Cambridge University Press; 1988.
3. Baars BJ, Banks WP, Newman J, editors. Essential sources in the scientific study of consciousness. Cambridge: MIT Press; 2003.
4. Bedau M. Weak emergence. Philos Perspect. 1997;11:375–99.
5. Bishop R, Atmanspacher H. Contextual emergence in the description of properties. Found Phys. 2006;36:1753–77.
6. Boly M, Balteau E, Schnakers C, Degueldre C, Moonen G, Luxen A, et al. Baseline brain activity fluctuations predict somatosensory perception in humans. Proc Natl Acad Sci USA. 2007;104(29):12187–92.
7. Bongard J, Zykov V, Lipson H. Resilient machines through continuous self-modeling. Science. 2006;314(5802):1118–21.
8. Bosse T, Jonker CM, Treur J. Formalization of Damasio's theory of emotion, feeling and core consciousness. Conscious Cogn. 2008;17(1):94–113.
9. Braitenberg V. Vehicles: experiments in synthetic psychology. Cambridge: MIT Press; 1984.
10. Bressler SL, Kelso JA. Cortical coordination dynamics and cognition. Trends Cogn Sci. 2001;5(1):26–36.
11. Burgess N. Spatial cognition and the brain. Ann N Y Acad Sci. 2008;1124:77–97.
12. Chalmers DJ. Strong and weak emergence. In: Clayton P, Davies P, editors. The re-emergence of emergence. Oxford: Oxford University Press; 2006.
13. Chang H. Inventing temperature: measurement and scientific progress. New York: Oxford University Press; 2004.
14. Cisek P. Cortical mechanisms of action selection: the affordance competition hypothesis. Philos Trans R Soc Lond B Biol Sci. 2007;362(1485):1585–99.
15. Clowes RW, Seth AK. Axioms, properties and criteria: roles for synthesis in the science of consciousness. Artif Intell Med. 2008;44:93–104.
16. Cosmelli D, Lachaux J-P, Thompson E. Neurodynamics of consciousness. In: Zelazo PD, Moscovitch M, Thompson E, editors. The cambridge handbook of consciousness. Cambridge: Cambridge University Press; 2007. p. 731–75.
17. Craig AD. How do you feel? Interoception: the sense of the physiological condition of the body. Nat Rev Neurosci. 2002;3(8):655–66.
18. Crick F, Koch C. Towards a neurobiological theory of consciousness. Semin Neurosci. 1990;2:263–75.
19. Critchley HD, Wiens S, Rotshtein P, Ohman A, Dolan RJ. Neural systems supporting interoceptive awareness. Nat Neurosci. 2004;7(2):189–95.
20. Cruse H. The evolution of cognition: a hypothesis. Cogn Sci. 2003;27:135–55.
21. Crutchfield J. The calculi of emergence: computation, dynamics, and induction. Physica D. 1994;75:11–54.
22. Damasio A. Descartes' error. London: MacMillan; 1994.
23. Damasio A. The feeling of what happens: body and emotion in the making of consciousness. Arlington Heights: Harvest Books; 2000.
24. Dehaene S, Sergent C, Changeux JP. A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc Natl Acad Sci USA. 2003; 100(14):8520–5.

25. Doya K. Modulators of decision making. Nat Neurosci. 2008;11(4):410–6.
26. Edelman GM. The remembered present. New York: Basic Books; 1989.
27. Edelman GM. Naturalizing consciousness: a theoretical framework. Proc Natl Acad Sci USA. 2003;100(9):5520–4.
28. Edelman DB, Baars BJ, Seth AK. Identifying the hallmarks of consciousness in non-mammalian species. Conscious Cogn. 2005;14(1):169–87.
29. Ehrsson HH. The experimental induction of out-of-body experiences. Science. 2007;317(5841):1048.
30. Engel AK, Singer W. Temporal binding and the neural correlates of sensory awareness. Trends Cogn Sci. 2001;5(1):16–25.
31. Fellous J-M, Arbib MA, editors. Who needs emotions? The brain meets the robot. Oxford: Oxford University Press; 2005.
32. Franklin S, Graesser A. A software agent model of consciousness. Conscious Cogn. 1999;8(3):285–301.
33. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, et al. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. Science. 2008;319(5867):1215–20.
34. Gould SJ, Lewontin RC. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc Lond B Biol Sci. 1979;205(1161):581–98.
35. Grandjean D, Sander D, Scherer KR. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Conscious Cogn. 2008;17(2):484–95.
36. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 1969;37:424–38.
37. Grossberg S, Gutowski WE. Neural dynamics of decision making under risk: affective balance and cognitive-emotional interactions. Psychol Rev. 1987;94(3):300–18.
38. Grush R. The emulation theory of representation: motor control, imagery, and perception. Behav Brain Sci. 2004;27(3):377–96; discussion 396–442.
39. Haggard P. Human volition: towards a neuroscience of will. Nat Rev Neurosci. 2008;9(12):934–46.
40. Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey CJ, Wedeen VJ, et al. Mapping the structural core of human cerebral cortex. PLoS Biol. 2008;6(7):e159.
41. Haugeland J. Artificial intelligence: the very idea. Cambridge: MIT Press; 1985.
42. Hesslow G. Conscious thought as simulation of behaviour and perception. Trends Cogn Sci. 2002;6(6):242–7.
43. Hesslow G, Jirenhed D-A. The inner world of a simple robot. J Conscious Stud. 2007;14:85–96.
44. Holland O. Editorial introduction. J Conscious Stud. 2003;10(4/5):1–6.
45. Holland O. A strongly embodied approach to machine consciousness. J Conscious Stud. 2007;14:97–110.
46. Hussain A. (this volume). Editorial introduction.
47. Izhikevich EM, Edelman GM. Large-scale model of mammalian thalamocortical systems. Proc Natl Acad Sci USA. 2008;105(9):3593–8.
48. James W. Does consciousness exist? J Philos Psychol Sci Methods. 1904;1:477–91.
49. Kim J. Emergence: core ideas and issues. Synthese. 2006;151:547–59.
50. Koch C. The quest for consciousness: a neurobiological approach. Englewood: Roberts and co; 2004.
51. Koechlin E, Hyafil A. Anterior prefrontal function and the limits of human decision-making. Science. 2007;318(5850):594–8.
52. Lambie JA, Marcel AJ. Consciousness and the varieties of emotion experience: a theoretical framework. Psychol Rev. 2002;109(2):219–59.
53. Lamme V. Towards a true neural stance on consciousness. Trends Cogn Sci. 2006;10(11):494–501.
54. Laureys S, Pellas F, Van Eeckhout P, Ghorbel S, Schnakers C, Perrin F, et al. The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? Prog Brain Res. 2005;150:495–511.
55. Lenggenhager B, Tadi T, Metzinger T, Blanke O. Video ergo sum: manipulating bodily self-consciousness. Science. 2007;317(5841):1096–9.
56. Libet B. Unconscious cerebral initiative and the role of conscious will in voluntary action. Behav Brain Sci. 1985;8:529–66.
57. Mandik P. Phenomenal consciousness and the allocentric-egocentric interface. In: Buccheri R, editor. Endophysics, time, quantum and the subjective. New York: World Scientific Publishing Co; 2005.
58. Mason MF, Norton MI, Van Horn JD, Wegner DM, Grafton ST, Macrae CN. Wandering minds: the default network and stimulus-independent thought. Science. 2007;315(5810):393–5.
59. Maturana H, Varela F. Autopoiesis and cognition: the realization of the living, vol. 42. Dordrecht: D. Reidel; 1980.
60. Mehta B, Schaal S. Forward models in visuomotor control. J Neurophysiol. 2002;88(2):942–53.
61. Merker B. The liabilities of mobility: a selection pressure for the transition to consciousness in animal evolution. Conscious Cogn. 2005;14(1):89–114.
62. Metzinger T. Being no-one. Cambridge: MIT Press; 2003.
63. Metzinger T. Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples. Prog Brain Res. 2008;168:218–45.
64. Nagel T. What is it like to be a bat? Philos Rev. 1974;83:435–50.
65. Northoff G, Panksepp J. The trans-species concept of self and the subcortical-cortical midline system. Trends Cogn Sci. 2008;12(7):259–64.
66. O'Regan JK, Noe A. A sensorimotor account of vision and visual consciousness. Behav Brain Sci. 2001;24(5):939–73; discussion 973–1031.
67. Panksepp J. Affective consciousness: core emotional feelings in animals and humans. Conscious Cogn. 2005;14(1):30–80.
68. Pessoa L. On the relationship between emotion and cognition. Nat Rev Neurosci. 2008;9(2):148–58.
69. Phillips ML, Medford N, Senior C, Bullmore ET, Suckling J, Brammer MJ, et al. Depersonalization disorder: thinking without feeling. Psychiatry Res. 2001;108(3):145–60.
70. Prescott TJ, Bryson JJ, Seth AK. Modelling natural action selection (edited special issue). Philos Trans R Soc Lond B Biol Sci. 2007;362(1485):1519–721.
71. Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. A default mode of brain function. Proc Natl Acad Sci USA. 2001;98(2):676–82.
72. Ramachandran VS, Rogers-Ramachandran D. Synaesthesia in phantom limbs induced with mirrors. Proc Biol Sci. 1996;263(1369):377–86.
73. Rees G, Kreiman G, Koch C. Neural correlates of consciousness in humans. Nat Rev Neurosci. 2002;3(4):261–70.
74. Revonsuo A. Inner presence: consciousness as a biological phenomenon. Cambridge: MIT Press; 2005.
75. Searle J. Minds, brains, and programs. Behav Brain Sci. 1980;3:417–57.
76. Seth AK. Causal connectivity analysis of evolved neural networks during behavior. Network: Comput Neural Syst. 2005;16(1):35–55.
77. Seth AK. Causal networks in simulated neural systems. Cogn Neurodyn. 2008;2:49–64.
78. Seth AK. Measuring emergence via nonlinear Granger causality. In: Bullock S, Watson R, Noble J, Bedau M, editors. Artificial

life XI: proceedings of the 11th international conference on the simulation and synthesis of living systems. Cambridge: MIT Press; 2008. p. 41–9.
79. Seth AK. Functions of consciousness. In: Banks WP, editor. Elsevier encyclopedia of consciousness. Amsterdam: Elsevier (in press).
80. Seth AK, Edelman GM. Environment and behavior influence the complexity of evolved neural networks. Adapt Behav. 2004;12(1):5–20.
81. Seth AK, Edelman, GM. Consciousness and complexity. In: Meyer B, editor. Springer encyclopedia of complexity and systems science. Berlin: Springer (in press).
82. Seth AK, Izhikevich E, Reeke GN, Edelman GM. Theories and measures of consciousness: an extended framework. Proc Natl Acad Sci USA. 2006;103(28):10799–804.
83. Seth AK, Dienes Z, Cleeremans A, Overgaard M, Pessoa L. Measuring consciousness: relating behavioural and neurophysiological approaches. Trends Cogn Sci. 2008;12(8):314–21.
84. Shadlen MN, Gold JI. The neurophysiology of decision-making as a window on cognition. In: Gazzaniga MS, editor. The cognitive neurosciences. 3rd ed. Cambridge: MIT Press; 2004. p. 1229–41.
85. Shalizi C, Moore C. What is a macrostate? Subjective observations and objective dynamics. 2006. http://arxiv.org/abs/cond-mat/0303625.
86. Shanahan M. A cognitive architecture that combines internal simulation with a global workspace. Conscious Cogn. 2006; 15(2):433–49.
87. Shanahan M. Dynamical complexity in small-world networks of spiking neurons. Phys Rev E Stat Nonlin Soft Matter Phys. 2008;78(4 Pt 1):041924.
88. Sporns O, Lungarella M. Evolving coordinated behavior by maximizing information structure. In: Rocha L, Yaeger L, Bedau M, Floreano D, Goldstone RL, Vespigniani A, editors. Artificial life X: proceedings of the 10th international conference on the simulation and synthesis of living systems. Cambridge: MIT Press; 2006. p. 322–9.
89. Sporns O, Tononi G, Edelman GM. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. Cereb Cortex. 2000;10:127–41.
90. Thagard P, Aubie B. Emotional consciousness: a neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. Conscious Cogn. 2008;17(3): 811–34.
91. Thompson E. Life and mind: from autopoeisis to neurophenomenology: a tribute to Francisco Varela. Phenomenol Cogn Sci. 2004;3:381–98.
92. Thompson E, Varela FJ. Radical embodiment: neural dynamics and consciousness. Trends Cogn Sci. 2001;5(10):418–25.
93. Tononi G. An information integration theory of consciousness. BMC Neurosci. 2004;5(1):42.
94. Tononi G, Edelman GM. Consciousness and complexity. Science. 1998;282(5395):1846–51.
95. Tononi G, Koch C. The neural correlates of consciousness: an update. Ann N Y Acad Sci. 2008;1124:239–61.
96. Tononi G, Sporns O. Measuring information integration. BMC Neurosci. 2003;4(1):31.
97. Tononi G, Sporns O, Edelman GM. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc Natl Acad Sci USA. 1994;91(11):5033–7.
98. Tsuchiya N, Adolphs R. Emotion and consciousness. Trends Cogn Sci. 2007;11(4):158–67.
99. Vallar G, Ronchi R. Somatoparaphrenia: a body delusion. A review of the neuropsychological literature. Exp Brain Res. 2008;192(3):533–51.
100. Varela FJ. Patterns of life: intertwining identity and cognition. Brain Cogn. 1997;34(1):72–87.
101. Wagar BM, Thagard P. Spiking phineas gage: a neurocomputational theory of cognitive-affective integration in decision making. Psychol Rev. 2004;111(1):67–79.
102. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440–2.
103. Wegner D. The illusion of conscious will. Cambidge: MIT Press; 2002.
104. Werner G. Metastability, criticality and phase transitions in brain and its models. Biosystems. 2007;90(2):496–508.
105. Wolpert DM, Kawato M. Multiple paired forward and inverse models for motor control. Neural Netw. 1998;11(7–8):1317–29.
106. Yaeger L, Sporns O. Evolution of neural structure and complexity in a computational ecology. In: Rocha L, Yaeger L, Bedau M, Floreano D, Goldstone RL, Vespigniani A, editors. Artificial life X: proceedings of the 10th international conference on the simulation and synthesis of living systems. Cambridge: MIT Press; 2006, p. 330–6.
107. Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005;46(4):681–92.
108. Zeman A. What in the world is consciousness. Prog Brain Res. 2005;150:1–10.
109. Ziemke T. The embodied self—theories, hunches, and robot models. J Conscious Stud. 2007;14:167–79.
110. Ziemke T, Jirenhed D-A, Hesslow G. Internal simulation of perception: a minimal neurorobotic model. Neurocomputing. 2005;68:85–104.