Opinion

Cell PRESS

Interoceptive inference, emotion, and the embodied self

Anil K. Seth

Sackler Centre for Consciousness Science and School of Engineering and Informatics, University of Sussex, Brighton, BN1 9QJ, UK

The concept of the brain as a prediction machine has enjoyed a resurgence in the context of the Bayesian brain and predictive coding approaches within cognitive science. To date, this perspective has been applied primarily to exteroceptive perception (e.g., vision, audition), and action. Here, I describe a predictive, inferential perspective on interoception: 'interoceptive inference' conceives of subjective feeling states (emotions) as arising from actively-inferred generative (predictive) models of the causes of interoceptive afferents. The model generalizes 'appraisal' theories that view emotions as emerging from cognitive evaluations of physiological changes, and it sheds new light on the neurocognitive mechanisms that underlie the experience of body ownership and conscious selfhood in health and in neuropsychiatric illness.

The predictive brain, body, and self

The view that prediction and error correction provide fundamental principles for understanding brain operation is gaining increasing traction within the cognitive and brain sciences. In the renascent guise of 'predictive coding' (PC – see Glossary) or 'predictive processing', perceptual content is seen as resulting from probabilistic, knowledge-driven inference on the external causes of sensory signals [1–4]. Here, this framework is applied to interoception, the sense of the internal physiological condition of the body [5], in order to elaborate a model of emotion as 'interoceptive inference' [6–9]. Interoceptive predictive coding – equivalently here, interoceptive inference – is hypothesised to engage an extended autonomic neural substrate with emphasis on the anterior insular cortex (AIC) as a comparator. This view extends alternative frameworks for understanding emotion [10–14] by proposing that emotional content is generated by active 'top-down' inference of the causes of interoceptive signals in a predictive coding context. It also extends previous models of insular cortex as supporting error-based learning of feeling states and uncertainty [15] and as responding to interoceptive mismatches that underlie anxiety [16].

Corresponding author: Seth, A.K. (a.k.seth@sussex.ac.uk).

Keywords: interoception; predictive coding; emotion; experience of body ownership; rubber hand illusion; active inference.

1364-6613

 http://dx.doi.org/10.1016/j.tics.2013.09.007

Representations of physiological conditions have frequently been associated with basic pre-reflective forms of selfhood [11], with the AIC occupying a central role on some views [13]. Selfhood is a constellation concept that involves not only representation and control of physiological homeostasis, but also the experience of owning and identifying with a particular body, the emergence of a first-person perspective, intention and agency, and metacognitive aspects that relate to the subjective 'I' and the narrative linking of episodic memories over time [17,18]. Here, I apply the framework of interoceptive inference to the experience of body ownership (EBO) as a central aspect of selfhood, proposing on the basis of recent data [19,20] that EBO is shaped by predictive multisensory integration

Glossary

Active inference: an extension of PC (and part of the free energy principle), which says that agents can suppress prediction errors by performing actions to bring about sensory states in line with predictions.
Augmented reality: a technique in which virtual images can be combined with real-world real-time visual input to create hybrid perceptual scenes that are usually presented to a subject via a head-mounted display.
Appraisal theories of emotion: a long-standing tradition, dating back to James (but not Lange), according to which emotions depend on cognitive interpretations of physiological changes.
Emotion: an affective state with psychological, experiential, behavioral, and visceral components. Emotional awareness refers to conscious awareness of an emotional state.
Experience of body ownership (EBO): the experience of certain parts of the world as belonging to one's body. EBO can be distinguished into that related to body parts (e.g., a hand) and a global sense of identification with a whole body.
Free energy principle: a generalization of PC according to which organisms minimize an upper bound on the entropy of sensory signals (the free energy). Under specific assumptions, free energy translates to prediction error.
Generative model: a probabilistic model that links (hidden) causes and data, usually specified in terms of likelihoods (of observing some data given their causes) and priors (on these causes). Generative models can be used to generate inputs in the absence of external stimulation.
Interoception: the sense of the internal physiological condition of the body.
Interoceptive sensitivity: a characterological trait that reflects individual sensitivity to interoceptive signals, usually operationalized via heartbeat detection tasks.
Predictive coding (PC): a data processing strategy whereby signals are represented by generative models. PC is typically implemented by functional architectures in which top-down signals convey predictions and bottom-up signals convey prediction errors.
Rubber hand illusion (RHI): a classic experiment in which the experience of body ownership is manipulated via perceptual correlations such that a fake (i.e., rubber) hand is experienced as part of a subject's body.
Selfhood: the experience of being a distinct, holistic entity, capable of global self-control and attention, possessing a body and a location in space and time [64]. Selfhood operates on multiple levels – from basic physiological representations to metacognitive and narrative aspects.
Subjective feeling states: consciously experienced emotional states that underlie emotional awareness.
Von Economo neurons (VENs): long-range projection neurons found selectively in hominid primates and certain other species. VENs are found preferentially in the AIC and ACC.

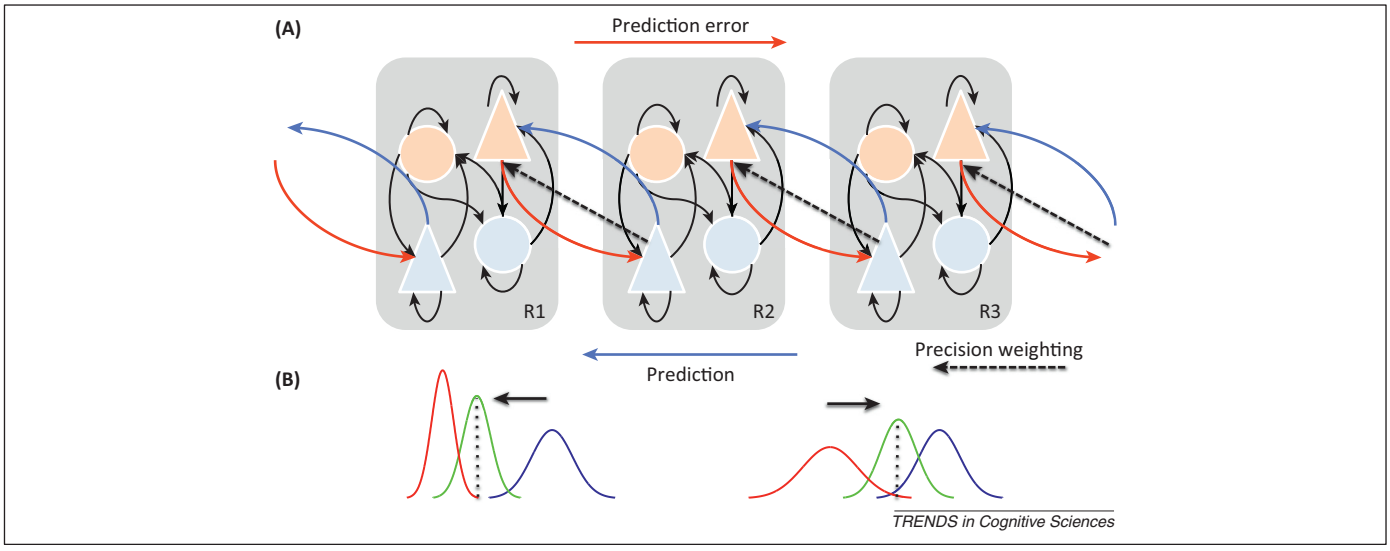

Figure 1. Functional architecture of predictive coding. **(A)** A schematic of hierarchical predictive coding across three cortical regions; the 'lowest' on the left (R1) and the 'highest' on the right (R3). Bottom-up projections (red) originate from 'error units' (light orange) in superficial cortical layers and terminate on 'state units' (light blue) in the deep (infragranular) layers of their targets, whereas top-down projections (dark blue) that convey predictions originate in deep layers and project to superficial layers of their targets. Prediction errors are associated with precisions (inverse variances), which determine the relative influence of bottom-up and top-down signal flow. Top-down precision weighting (dashed lines) regulates the post-synaptic gain of prediction-error projection neurons possibly by neuromodulation. Triangles represent pyramidal (projection) neurons; circles represent inhibitory interneurons. Solid black lines depict local circuit interactions wherein descending predictions are resolved with ascending prediction errors. **(B)** The influence of precisions on Bayesian inference and predictive coding. The curves represent probability distributions over the value of a sensory signal (*x*-axis). On the left, high precision-weighting of sensory signals (red) enhances their influence on the posterior (green) and expectation (dotted line) as compared to the prior (blue). On the right, low precision-weighting of sensory signals has the opposite effect on posteriors and expectations. Adapted from [6].

of self-related signals across interoceptive and exteroceptive domains.

Overall, the model described here provides a unified view of self-related processing relevant to emotional awareness and EBO, and carries implications for understanding specific neuropsychiatric disorders.

Predictive inference and perception

The concept of PC overturns classical notions of perception as a largely 'bottom-up' process of evidence accumulation or feature detection, proposing instead that perceptual content is specified by top-down predictive signals that emerge from hierarchically organized generative models of the causes of sensory signals. According to PC, the brain is continuously attempting to minimize the discrepancy or 'prediction error' between its inputs and its emerging models of the causes of these inputs via neural computations approximating Bayesian inference (Figure 1 and Box 1). Importantly, prediction errors can be minimized either by updating generative models (perceptual inference and learning; changing the model to fit the world) or by performing actions to bring about sensory states in line with predictions (active inference; changing the world to fit the model). In most incarnations these processes are assumed to unfold continuously and simultaneously, underlining a deep continuity between perception and

Box 1. PC, free energy, and active inference

PC has a long history, originating with the insights of von Helmholtz and reaching recent prominence in the 'Bayesian brain' hypothesis [1,4]. The idea is that, in order to support adaptive responses, the brain must discover information about the likely causes of sensory signals (i.e., perception) without direct access to these causes, using only information in the flux of sensory signals themselves [2]. According to PC, this is accomplished via probabilistic inference on the causes of sensory signals, computed according to Bayesian principles. This means estimating the probable causes of data (the posterior) given observed conditional probabilities (likelihoods) and prior 'beliefs' about probable causes. This, in turn, means inducing a predictive or 'generative' model of the sensory data. Although exact Bayesian inference is computationally challenging and often intractable, a variety of approximate methods exist. Within neuroscience, these approximations have been elaborated in Friston's 'free energy principle' [2,65], which, following seminal work by Hinton and colleagues [66,67], shows how generative models can be induced from data by assuming that the brain minimizes a bound on the evidence for this data (the 'free energy', which under simplifying (Gaussian) assumptions is equivalent to prediction error). The generalization of Bayes theorem to a hierarchical context implies that posteriors at one level form the priors at one level lower, thus enabling priors to be induced from the data stream itself ('empirical' Bayes). Applied to cortical networks, PC interprets bottom-up signals as conveying prediction errors and top-down signals as conveying predictions (Figure 1). Although unequivocal neural evidence for PC is still lacking, a growing body of supportive data details how perceptual content – and underlying neural responses – can be shaped by pre-stimulus expectations [1,4,42].

Key to PC is the minimization of prediction error across hierarchical levels. This can be accomplished either by updating generative models to accommodate unexpected sensory signals or by performing actions to confirm sensory predictions (active inference, [21]). This duality underlines a strong continuity between perception, action, and imagination [68]. Also important is that prediction errors are associated with precisions (Figure 1), so that dynamic precision-weighting (for example by attention) can modulate the balance between top-down and bottom-up signal flow (e.g., low precision on error signals corresponds to high confidence in top-down prior beliefs). The present framework generalizes PC to interoception, proposing that affective states depend on active inference of interoceptive responses.

action. Prediction errors are associated with 'precisions', which determine their influence on subsequent hierarchical processing. For example, precision weighting (possibly implemented by post-synaptic gain modulation of prediction error units) can modulate the extent to which prediction errors are resolved by updating generative models or by performing actions [21]. This leads to an interpretation of attention as the optimization of precision weighting, balancing the relative influence of prediction errors and prior expectations on perceptual inference [2].

PC has been elaborated principally in the context of exteroception; that is, to predictive modelling of external states of the world. However, one of the most relevant features of the world for a particular organism is the organism itself [11,22]. This reflects a long-standing notion that mental representations of selfhood are ultimately grounded in representations of the body, with the internal physiological milieu providing a primary reference – a 'material me' [13] – that supports adaptive interaction with the environment. From a PC perspective, this implies that an organism should maintain well-adapted predictive models of its own physical body (its position, morphology, etc.) and of its internal physiological condition. This entails inducing generative models of the causes of those signals 'most likely to be me' [22] across interoceptive and exteroceptive domains, a framework that views emotion as 'interoceptive inference' and provides a unifying mechanism for self-representation at multiple levels, including perhaps especially those related to EBO.

Interoceptive inference and emotion

Interoceptive concepts of emotion were crystallized by James and Lange [10], who argued that emotions arise from perceptions of changes in the body. This approach evolved into 'appraisal' theories, which recognise that explicit cognitions and beliefs about the causes of physiological changes influence subjective feeling states and emotional behaviour [23]. Schachter and Singer [24] famously demonstrated that injections of adrenaline, proximally causing a state of physiological arousal, would give rise to either anger or elation depending on the context (an irritated or elated confederate). This observation was formalized in their 'two factor' theory, in which emotional experience is determined by the combination of physiological change and cognitive appraisal, that is, emotion as interpreted bodily arousal (see [25] for a precursor). More than a century after James and Lange, there is now a consensus that emotions are psychological states that encompass behavioural, experiential, and visceral changes [23,26–28]. This attitude underpins several contemporary frameworks for understanding emotion and its relation to cognition and self [11–13], discussed further below.

Despite the above insights, interoception has remained generally understood along feed-forward lines, similar to classical evidence accumulation theories of exteroception [23,28]. This assumption is however challenged by evidence of substantial cross-talk between levels of viscerosensory representation, including top-down cortical and behavioural influences to brainstem and spinal centres [26]. Informed by this emerging picture, I suggest that the role of interoception in shaping emotion and selfhood can be productively understood through the lens of PC. In this view, interoceptive inference involves hierarchically cascading top-down interoceptive predictions that counterflow with bottom-up interoceptive prediction errors. Subjective feeling states – experienced emotions – are hypothesized to depend on the integrated content of these predictive representations across multiple levels [6].

Following PC principles, interoceptive prediction errors can be suppressed both by modifying predictions and by transcribing these predictions into reference points for autonomic reflexes that regulate physiological homeostasis, as recently suggested by Gu and colleagues [9]. This role for active inference, which extends previous presentations of this model [6–8], directly parallels PC formulations for motor control (e.g. [29]) which highlight descending corticospinal signals as instantiating proprioceptive predictions that engage classical motor reflexes. Precisions play a key role here: descending predictions can engage motor or autonomic reflexes only if the corresponding error signals have diminished impact on hierarchical processing via transiently low precision weighting, which corresponds to decreased attention to these error signals. Without this transient modulation, precise prediction errors would lead to revision of predictions rather than to action [29]. This implies that active interoceptive inference depends on the selective attenuation of attention to interoceptive prediction errors.

Interoceptive predictions arise from multiple hierarchical levels, with higher levels integrating interoceptive, proprioceptive and exteroceptive cues in formulating descending predictions. These multimodal predictions underwrite emotional responses to exteroceptive cues (which may include socially salient signals, see later). In short, interoceptive predictive coding (inference) proposes that emotional content is determined by active inference on the likely internal and external causes of changes in the physiological condition of the body (Figure 2).

Evidence for interoceptive inference

Although there is not yet any direct confirmatory evidence for interoceptive inference (as for PC generally, see [1]), supportive data are steadily accumulating. Much of these data rest on assuming a central role for the anterior insular cortex (AIC), operating within a rich functional network [30], both as a comparator that registers top-down predictions against bottom-up prediction errors and as a source of anticipatory visceromotor control [6–9,15,16]. Structurally, the AIC is ideally placed both to detect and to cause changes in physiological condition, and to integrate interoceptive and exteroceptive signals; functionally, it instantiates interoceptive representations accessible to conscious awareness and is associated with processes that involve visceral representation, interoception, and emotional awareness relevant to selfhood (Box 2).

Several functional MRI (fMRI) studies have shown anticipatory and prediction error responses within the AIC. For example, the AIC is activated by anticipation of painful [31] and affect-laden touch [32], and AIC responses encode both predicted pain and pain prediction error within a single task [33]. Similarly, in a gambling task, the AIC has been shown to encode both predicted risk

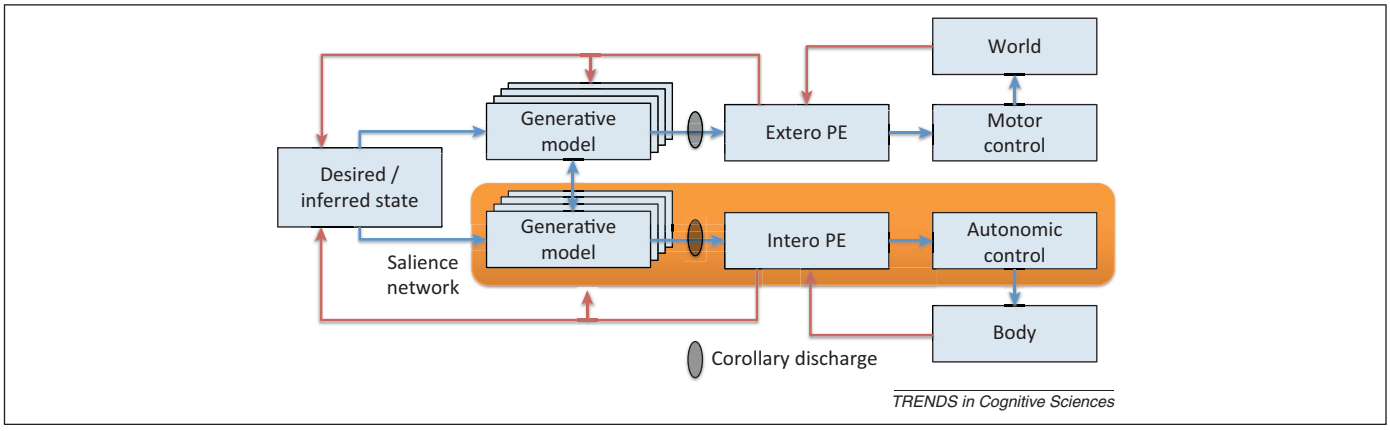

Figure 2. A model of interoceptive inference. In the model, emotional responses depend on continually-updated predictions of the causes of interoceptive input. Starting with a desired or inferred physiological state (which is itself subject to update based on higher-level motivational and goal-directed factors), generative models are engaged which predict interoceptive (and exteroceptive) signals via corollary discharge. Applying active inference, prediction errors (PEs) are transcribed into actions via engagement of classical reflex arcs (motor control) and autonomic reflexes (autonomic control). The resulting prediction error signals are used to update the (functionally coupled) generative models and the inferred/desired state of the organism. (At high hierarchical levels these generative models merge into a single multimodal model.) Interoceptive predictions are proposed to be generated, compared, and updated within a salience network (orange shading) anchored on the anterior insular and anterior cingulate cortices (AIC, ACC) that engage brainstem regions as targets for visceromotor control and relays of afferent interoceptive signals. Sympathetic and parasympathetic outflow from the AIC and ACC are in the form of interoceptive predictions that enslave autonomic reflexes (e.g., heart/respiratory rate, smooth muscle behaviour), just as proprioceptive predictions enslave classical motor reflexes in PC formulations of motor control [9]. This process depends on the transient attenuation of the precision of interoceptive (and proprioceptive) PE signals. Blue (red) arrows signify top-down (bottom-up) connections respectively.

and risk prediction error [34]. Holle *et al.* found AIC activation on viewing movies of people scratching, with the degree of activation correlating with the subjective level of reported itchiness [35]. Here, AIC responses plausibly reflect a conflict between predicted and actual interoceptive responses given anticipated itch, which suggests an engagement of AIC in predictive inference related to mirror-system activity and empathy [36]. Using the neuroeconomic 'ultimatum game', Xiang and colleagues showed that subjects used Bayesian rules to update expected feelings given economic outcomes, and that feeling prediction errors (when measured using variance) scaled parametrically with AIC activation [37]. These studies support the idea that interoceptive inference extends into social contexts to explain emotional behaviour. For example, the expression of an emotional state (e.g., anger) elicits behavioural responses from others, the detection of which could serve to confirm predictions of interoceptive condition. This provides a nice link to predictive models of social interaction [38,39].

Supportive evidence is also provided by studies that manipulate interoceptive feedback. Early evidence showed that false cardiac feedback enhanced subjective ratings of emotive stimuli [40]. This was later confirmed by Gray and colleagues, who found that false cardiac feedback led to increased fMRI responses in the right AIC and that enhanced AIC activity during false feedback correlated with increased emotional salience attributed to previously unthreatening stimuli [41]. Further evidence showing the influence of false interoceptive feedback on the experience of body ownership is described below. Taken together with the structural and functional attributes of the AIC (Box 2),

Box 2. The anterior insula cortex and interoception

The human insular cortex is found bilaterally beneath the temporal and frontal lobes, enjoying widespread bidirectional connectivity to parietal, frontal, and limbic regions [30]. Interoceptive pathways have their primary cortical representation within the insula, which contains a viscerotopic map [69]. A posterior-to-anterior gradient has been proposed, whereby posterior regions support primary (objective) mappings of interoceptive signals, whereas the anterior insula (AIC) supports secondary re-representations that underlie subjective access [9,12,13] and integration of interoceptive, motivational, and exteroceptive signals [6,70]. The AIC is closely connected structurally and functionally with the anterior cingulate cortex (ACC) as part of a cortical 'salience network' [71], though they can be functionally differentiated [70], so that the AIC can be considered a 'limbic sensory area' and the ACC a 'limbic motor area' [13,70]. These areas together engage subcortical regions, such as the periaqueductal gray matter and parabrachial nucleus, as targets for visceromotor control and relays for viscerosensory afferents, as well as many other areas related to self and emotion, including the amygdala, nucleus accumbens, and orbitofrontal cortex. Interestingly, the AIC contains a high density of von Economo neurons (VENs) [72], which have been associated frequently though circumstantially with conscious awareness and selfhood [8,13,73]. VENs are large projection neurons well suited for rapid long-range information integration and so may underlie the efficient registration of interoceptive prediction and prediction error signals. In PC, top-down predictions are suggested to originate in infragranular pyramidal cells (Figure 1), consistent with the layer V predominance of VENs in the AIC [8,9,73].

The AIC is engaged in a wide range of processes that share as common factors visceral representation, interoception, and emotional awareness [6,9,13,26]. Importantly, the AIC (particularly on the right) appears to support conscious access to both interoceptive information – as reflected in subjective emotional states (though see [43,74]) – and to associated representations of how encoded objects and contexts relate to the biological self [15,75]. Distinct axes of interoceptive information map onto specific insula sub-regions giving rise to dissociable emotional feeling states [76]. Also, individual differences in interoceptive sensitivity (IS), as measured by heartbeat detection, are predicted by AIC activation and morphometry, and are also associated with reported emotional symptoms [12] and susceptibility to illusions of selfhood (see main text).

These structural and functional attributes, when considered together with evidence for prediction error processing, strongly suggest the AIC as a locus for comparator mechanisms that underlie interoceptive inference, in turn supporting emotional awareness and consciously accessible integrated self-representations.

these data provide indirect support for interoceptive inference and the central involvement of the AIC in prediction error registration underlying subjective feeling states. Additional experiments, perhaps based on designs utilized in exteroceptive contexts (e.g., [42]), are needed to test more directly interoceptive inference and the proposed role of the AIC.

Relation to other models of emotion and insula function

The present model is related to several contemporary frameworks for understanding emotion and self, including prominently those of Damasio [11], Craig [13], and Critchley [12]. Each of these sees selfhood as grounded in representations of physiological condition, with the AIC emphasised in some [12,13], but not all [43]. Crucially, none identify emotional states with top-down inference of the causes of interoceptive signals, as argued here (and as recently taken up by others [9]). Rather, they emphasize a continuous, dynamic, but largely bottom-up interoceptive representational hierarchy that interacts with other perceptions to motivate behaviour. Perhaps more closely related is the suggestion that the AIC is involved in error-based learning of feeling states and uncertainty [15]; here, the notion of 'prediction error' is expressed in terms of change detection and salience rather than through mechanisms of predictive coding. Also relevant are models of insular dysfunction during anxiety [16] and psychosis [44] that hypothesize the existence of interoceptive prediction errors. However, although these models integrate abundant evidence compatible with a role for predictive interoception underlying emotion and self, the core notion of interoceptive inference is not elaborated.

The predictive self and the experience of body ownership

A predictive model of selfhood must extend beyond subjective feelings to integrate interoceptive and exteroceptive signals across multiple levels of self-representation. Particularly significant is the representation of those parts of the world perceived as belonging to one's own body, supporting EBO [45], and which plausibly are closely tied to interoception in having the body as a referent. Experiments such as the 'rubber hand illusion' (RHI) attest to the plasticity of EBO: stroking of an artificial hand synchronously with a participant's real hand, while visual attention is focused on the artificial hand, leads the participant to experience the artificial hand as part of her own body [46]. Similar effects have been described for face perception [47] and whole-body ownership [48,49].

These manipulations of EBO have been explained by models of multisensory integration, which propose that conflicts between vision, touch, and proprioception are minimized by visual capture of visual and felt (tactile) events occurring in peri-personal space, on the basis of statistical correlations among sensory signals together with visual dominance [46,50]. This is compatible with PC inasmuch as minimization of prediction errors – such as those induced by multisensory conflicts during the RHI – will update the posterior probabilities and, over time, can induce changes in self-related priors [22]. Priors reflecting high-level representations, such as selfhood and body-ownership, are likely to operate at relatively abstract multisensory or amodal levels. Thus, statistical correlations among highly precision-weighted sensory signals (vision, touch) could overcome prediction errors in a different modality (proprioception), leading to a revised multisensory predictive model that minimizes the overall level of self-related precision-weighted prediction error by incorporating the fake hand as part of the self-representation. Moreover, the relative weighting of different information sources according to their reliability in this process is suggestive of the deployment of precision expectations which underlie optimal cue combination in a Bayesian scheme [51]. The engagement of predictive self-models is further supported by recent evidence showing that perceptual correlations are not necessary for the RHI; the expectation of correlated sensory input is sufficient [52]. Interestingly, EBO has also been associated with the integrity and activity of the AIC on the basis of lesion data and functional neuroimaging in healthy subjects [53].

Experiments have begun to address the impact of interoceptive processing on the modulation of EBO in paradigms like the RHI. One avenue is provided by individual differences in 'interoceptive sensitivity' (IS), which refers to a person's ability to detect their own interoceptive signals, operationalized by heartbeat detection tasks [54]. Tsakiris and colleagues reported that participants with lower IS are more susceptible to the RHI [55] and to the modulation of self–other boundaries in response to multisensory stimulation [56,57], possibly reflecting lower precision-weighting of interoceptive prediction errors. In addition, induction of the RHI leads to decreased temperature [58] and increased histamine reactivity of the real hand [59], whereas cooling the real hand increases susceptibility to the RHI [60]. Also, threats to the rubber hand during the RHI evoke enhanced skin conductance responses [61]. These effects are compatible with descending self-related predictions tuning reference points for autonomic reflexes from the perspective of active inference.

Two recent studies have addressed directly the role of multisensory integration across interoceptive and exteroceptive domains in shaping EBO. In the first, Suzuki *et al.* [19] combined augmented reality with physiological monitoring to implement a version of the RHI in which a virtual hand changed colour either in-time or out-of-time with an individual's heartbeat (Figure 3). The authors found that synchronous cardio-visual feedback enhanced the experience of ownership of the virtual hand, as measured both subjectively by questionnaire and objectively by previously validated measures of 'proprioceptive drift' – where the latter measures the perceived position of the real (hidden) hand by asking the participant to move a cursor to this location. (Note that these two measures do not always correlate [62].) The same setup enabled a replication of the standard RHI by contrasting synchronous versus asynchronous tactile-visual feedback. In a third condition, Suzuki *et al.* found that real-time remapping of finger movements to the virtual hand – supporting peri-hand visual-proprioceptive coherence – provided strong cues for EBO that overshadowed the influence of cardiac (interoceptive) prediction errors. A second study [20]

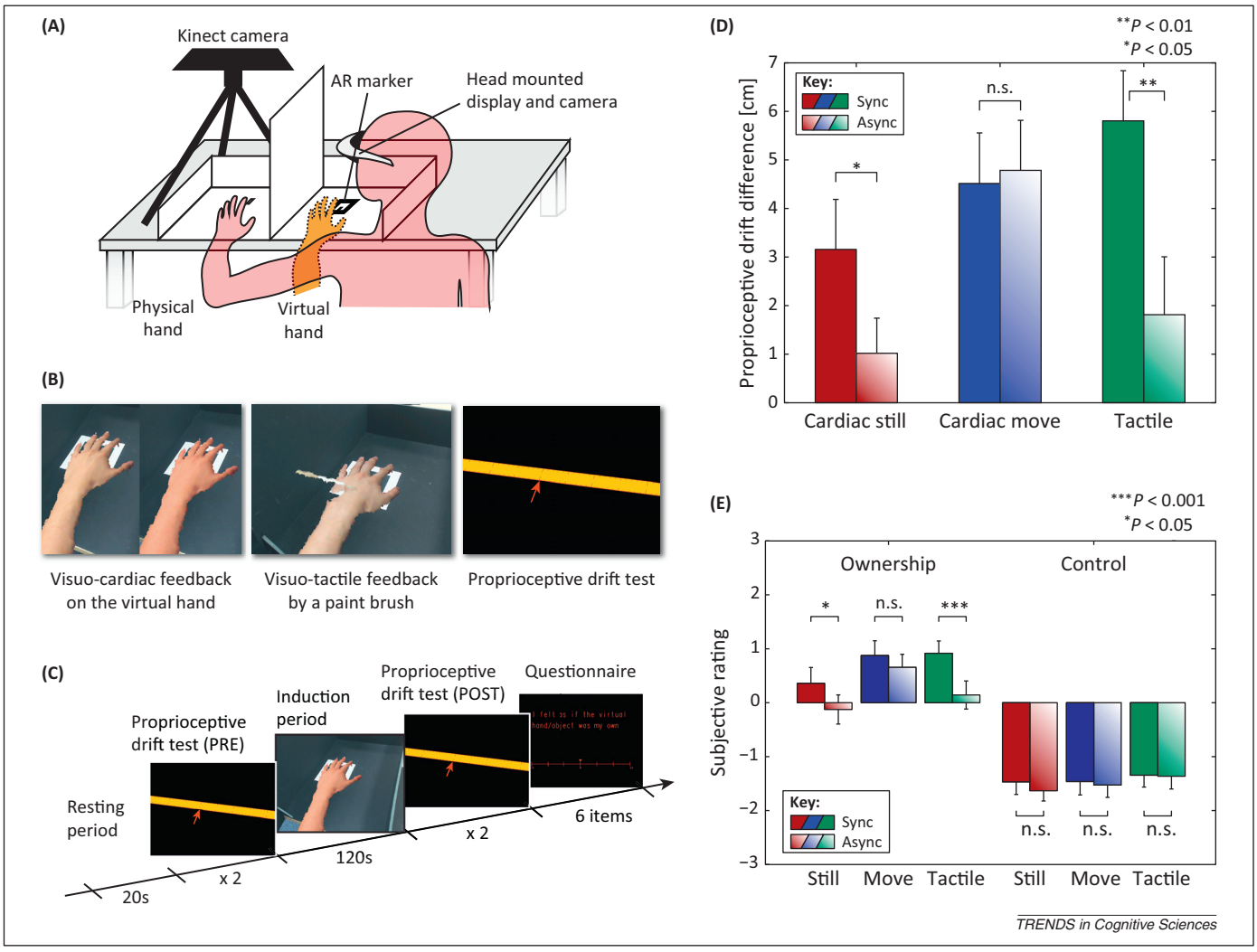

Figure 3. An interoceptive rubber-hand illusion. **(A)** Participants sat facing a desk so that their physical (left) hand was out of sight. A 3D model of the real hand was captured by Microsoft Kinect and used to generate a real-time virtual hand that was projected into the head-mounted display (HMD) at the location of the augmented-reality (AR) marker. Subjects wore a front-facing camera connected to the HMD, so they saw the camera image superimposed with the virtual hand. They also wore a pulse-oximeter to measure heartbeat timings and they used their right hand to make behavioural responses. **(B)** Cardio-visual feedback (left) was implemented by changing the colour of the virtual hand from its natural colour towards red and back, over 500 ms either synchronously or asynchronously with the heartbeat. Tactile feedback (middle) was given by a paintbrush, which was rendered into the AR environment. A 'proprioceptive drift' (PD) test (right), adapted for the AR environment, objectively measured perceived virtual hand position by implementing a virtual measure and cursor. **(C)** The experiment consisted of three blocks of four trials each. Each trial consisted of two PD tests flanking an induction period, during which either cardio-visual or tactile-visual feedback was provided (120 s). Each trial ended with a questionnaire presented in the HMD. **(D)** PD differences (PDD, post-induction minus pre-induction) were significantly larger for synchronous versus asynchronous cardio-visual feedback in the 'cardiac still' (without finger movements), but not the 'cardiac move' condition (with finger movements). PDDs were also significantly larger for synchronous versus asynchronous tactile-visual feedback ('tactile' condition), replicating the classical RHI. Each bar shows the across-participant average and standard error. **(E)** Subjective questionnaire responses that probed experience of ownership showed the same pattern as PDDs, whereas control questions showed no effect of cardio-visual or tactile-visual synchrony. Adapted from [19].

applied similar cardio-visual feedback to a virtual body as seen from behind, thus testing the influence of such feedback on the 'full body illusion' [48,49]. Consistent with Suzuki *et al.*'s results, they found that cardio-visual synchrony increased identification with the virtual body, again when measured both objectively and subjectively.

These data suggest that statistical correlations between interoceptive (e.g., cardiac) and exteroceptive (e.g., visual) signals can lead to updating of predictive models of self-related signals through minimization of prediction error, just as may happen for purely exteroceptive multisensory conflicts in the classic RHI. Also, if predictive models are continually probed by control signals that attempt to confirm the currently dominant model via active inference [21], the framework naturally accommodates the phasic physiological changes that accompany RHI induction [58,59] when these changes are taken to reflect altered autonomic control.

Concluding remarks

This opinion article proposes that emotion and embodied selfhood are grounded in active inference of those signals most likely to be 'me' across interoceptive and exteroceptive domains. In humans, self-related predictive coding simultaneously engages multiple levels of self-representation, including physiological homeostasis, physical bodily integrity, morphology and position, and – more speculatively – the metacognitive and narrative 'I'. Subjective feeling states (emotional experiences) arise from active interoceptive inference, extending previous theories based on cognitive appraisal of perceived physiological changes [24] and contemporary frameworks that emphasize

Box 3. Interoceptive inference and psychopathology

A significant body of work now connects deficits in predictive inference with psychiatric disorders that affect self-representation [6,16,44]. For two decades, 'comparator' models of schizophrenia have suggested that disturbances of selfhood (e.g., delusions of control) reflect problems in distinguishing self-caused from externally-caused changes in sensory input [77]. In psychosis, false perceptions (hallucinations) and false beliefs (delusions) may arise from reshaping of top-down predictions in attempts to suppress aberrant and persistent (exteroceptive) prediction errors [78]. This view has been finessed in terms of abnormal encoding of the relative precision of priors and sensory evidence to account for a broad range of psychotic symptoms [79,80].

Considering interoceptive inference and the AIC as its likely brain basis further enhances the explanatory potential of this framework for psychopathology. Links between perceptions of bodily state and emotional and neuropsychiatric disorders are well established. For example, alexithymia (broadly, deficits in emotional awareness) is associated with failure to engage AIC [81] and with degeneration of AIC and of VENs (Box 2) in the context of fronto-temporal dementia [82,83]. Human VENs express proteins linked to schizophrenia (notably DISC-1) and VEN density (in the anterior cingulate) has been linked to illness duration and completed suicide in psychotic patients [84]. Neuroimaging studies have correlated (right) AIC activity and volume with individual interoceptive sensitivity (IS), which in turn predicts sensitivity to mood states, including anxiety [12,16]. Indeed, anxiety and psychosis have been specifically associated with mismatches between predicted and actual interoceptive states putatively computed within AIC [16,44]. The finding that low IS predicts susceptibility to the RHI [55] provides a bridge to understanding how disorders of body image and ownership may involve disrupted interoceptive inference.

Dissociative conditions, such as depersonalization and derealisation (DPD), involve disabling disturbances of selfhood, as reflected in a persistent sense of unreality ('as-if-ness') [85]. These conditions are associated with alexithymia and a general loss of 'emotional colour', suggestive of deficient interoceptive inference and consistent with observed hypoactivation of AIC (and ACC bilaterally) in DPD patients when viewing aversive images [86]. By analogy with models of psychosis [79,80], Seth *et al.* have suggested that DPD may arise from imprecise (as opposed to inaccurate) interoceptive predictions, as part of a model of conscious 'presence' [6]. The transition from DPD to full-blown delusion (e.g., Cotard's syndrome, in which patients believe they are dead [87]) may also reflect aberrant high-level inference as the result of attempting to explain away persistent interoceptive prediction errors.

bottom-up elaboration of interoceptive representations with perception and motivation [11,13,15]. The close interplay between interoceptive and exteroceptive inference implies that emotional responses are inevitably shaped by cognitive and exteroceptive context, and that perceptual scenes that evoke interoceptive predictions will always be affectively coloured. Although the detailed neuroanatomy that underlies interoceptive inference remains to be elucidated, accumulating evidence implicates the AIC as a key comparator mechanism sensitive to interoceptive prediction error signals, as informing visceromotor control, and as underpinning conscious access to emotional states (emotional awareness). A predictive self is further supported by emerging paradigms that combine virtual/augmented reality and physiological monitoring, where the data so far suggest that the experience of body ownership, a key aspect of selfhood, is modulated by predictive multisensory integration of precision-weighted interoceptive and exteroceptive signals.

This framework may have important implications for understanding psychiatric disturbances of selfhood and emotional awareness. Dysfunctions in interoceptive inference could underlie a range of pathologies, especially those that involve dissociative symptoms, such as anxiety (Box 3). A broad role for interoceptive inference is also consistent with recent evidence that shows interoceptive influences on cognition and perception. For example, word recognition memory is modulated by the timing of visual stimuli with respect to cardiac phase under restricted attention [63]. Overall, 'interoceptive inference' highlights common predictive mechanisms that underlie affect, self, and perception, emphasizes their integration in shaping self-related experience, and provides new experimental possibilities for elucidating the underlying processes (Box 4).

Box 4. Questions for future research

- Can (structurally or functionally) segregated populations of representation and error units be identified in the interoceptive system? There are some early indications of such segregation in perceptual systems [88].
- Do neurotransmitters like dopamine and oxytocin have a role in modulating the precision of interoceptive prediction errors [89]? How do these roles relate to similar roles in reward processing [2]?
- Are VENs engaged in communicating top-down interoceptive predictions from insular and cingulate cortices to subcortical and brainstem targets [8]?
- What are the functional and effective connectivity patterns that underlie interoceptive inference and the integration of interoceptive and exteroceptive signals relevant to conscious selfhood?
- What is the relation between predictive models of interoception and similar models of agency [90], another central aspect of selfhood?
- Can predictive models of selfhood accommodate more abstract modes of self-experience, such as the narrative self or 'I' that links episodic memories across time?
- Can interoceptive feedback provide new avenues for the treatment of emotional and neuropsychiatric disorders? Might there be a role for disrupted interoceptive inference in impairments of social emotional behaviour, such as autism?

Acknowledgements

This work was supported by ERC FP7 project CEEDS (FP7-ICT-258749), EPSRC Fellowship EP/G007543/1, and by the Dr Mortimer and Dame Theresa Sackler Foundation, which supports the work of the Sackler Centre for Consciousness Science. I am grateful to Sarah Garfinkel, Jakob Hohwy, Keisuke Suzuki (who also helped prepare Figure 3), Paul Verschure, members of the Sackler Centre, and my reviewers for comments which helped improve the manuscript. Special gratitude is due to Hugo Critchley and Manos Tsakiris for their generous insights which helped shape the ideas presented here.

References

1 Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204
2 Friston, K.J. (2009) The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301
3 Lee, T.S. and Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A: Opt. Image Sci. Vis.* 20, 1434–1448
4 Hohwy, J. (2013) *The Predictive Mind*, Oxford University Press
5 Craig, A.D. (2002) How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* 3, 655–666
6 Seth, A.K. *et al.* (2011) An interoceptive predictive coding model of conscious presence. *Front. Psychol.* 2, 395

7 Seth, A.K. and Critchley, H.D. (2013) Extending predictive processing to the body: Emotion as interoceptive inference. *Behav. Brain Sci.* 36, 227–228
8 Critchley, H. and Seth, A. (2012) Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron* 74, 423–426
9 Gu, X. *et al.* (2013) Anterior insular cortex and emotional awareness. *J. Comp. Neurol.* 521, 3371–3388
10 James, W. (1894) The physical basis of emotion. *Psychol. Rev.* 1, 516–529
11 Damasio, A. (2000) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, Harvest Books
12 Critchley, H.D. *et al.* (2004) Neural systems supporting interoceptive awareness. *Nat. Neurosci.* 7, 189–195
13 Craig, A.D. (2009) How do you feel – now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70
14 Damasio, A. and Carvalho, G.B. (2013) The nature of feelings: evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* 14, 143–152
15 Singer, T. *et al.* (2009) A common role of insula in feelings, empathy and uncertainty. *Trends Cogn. Sci.* 13, 334–340
16 Paulus, M.P. and Stein, M.B. (2006) An insular view of anxiety. *Biol. Psychiatry* 60, 383–387
17 Metzinger, T. (2003) *Being No-One*, MIT Press
18 Northoff, G. and Bermpohl, F. (2004) Cortical midline structures and the self. *Trends Cogn. Sci.* 8, 102–107
19 Suzuki, K. *et al.* (2013) Multisensory integration across interoceptive and exteroceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia* http://dx.doi.org/10.1016/j.neuropsychologia.2013.08.014
20 Aspell, J.E. *et al.* (2013) Turning body and self inside out: visualized heartbeats alter bodily self-consciousness and tactile perception. *Psychol. Sci.* http://dx.doi.org/10.1177/0956797613498395
21 Friston, K.J. *et al.* (2010) Action and behavior: a free-energy formulation. *Biol. Cybern.* 102, 227–260
22 Apps, M.A. and Tsakiris, M. (2013) The free-energy self: a predictive coding account of self-recognition. *Neurosci. Biobehav. Rev.* http://dx.doi.org/10.1016/j.neubiorev.2013.01.029
23 Gendron, M. and Barrett, L.F. (2009) reconstructing the past: a century of ideas about emotion in psychology. *Emot. Rev.* 1, 316–339
24 Schachter, S. and Singer, J.E. (1962) Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* 69, 379–399
25 Cantril, H. and Hunt, W.A. (1932) Emotional effects produced by the injection of adrenalin. *Am. J. Psychol.* 44, 300–307
26 Critchley, H.D. and Harrison, N.A. (2013) Visceral influences on brain and behavior. *Neuron* 77, 624–638
27 Lane, R.D. and Schwartz, G.E. (1987) Levels of emotional awareness: a cognitive-developmental theory and its application to psychopathology. *Am. J. Psychiatry* 144, 133–143
28 Dolan, R.J. (2002) Emotion, cognition, and behavior. *Science* 298, 1191–1194
29 Adams, R.A. *et al.* (2013) Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218, 611–643
30 Deen, B. *et al.* (2011) Three systems of insular functional connectivity identified with cluster analysis. *Cereb. Cortex* 21, 1498–1506
31 Ploghaus, A. *et al.* (1999) Dissociating pain from its anticipation in the human brain. *Science* 284, 1979–1981
32 Lovero, K.L. *et al.* (2009) Anterior insular cortex anticipates impending stimulus significance. *Neuroimage* 45, 976–983
33 Seymour, B. *et al.* (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429, 664–667
34 Preuschoff, K. *et al.* (2008) Human insula activation reflects risk prediction errors as well as risk. *J. Neurosci.* 28, 2745–2752
35 Holle, H. *et al.* (2012) Neural basis of contagious itch and why some people are more prone to it. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19816–19821
36 Singer, T. *et al.* (2004) Empathy for pain involves the affective but not sensory components of pain. *Science* 303, 1157–1162
37 Xiang, T. *et al.* (2013) Computational substrates of norms and their violations during social exchange. *J. Neurosci.* 33, 1099–1108
38 Frith, C.D. and Frith, U. (2012) Mechanisms of social cognition. *Annu. Rev. Psychol.* 63, 287–313
39 Wolpert, D.M. *et al.* (2003) A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 358, 593–602
40 Valins, S. (1966) Cognitive effects of false heart-rate feedback. *J. Pers. Soc. Psychol.* 4, 400–408
41 Gray, M.A. *et al.* (2007) Modulation of emotional appraisal by false physiological feedback during fMRI. *PLoS ONE* 2, e546
42 Egner, T. *et al.* (2008) Neural repetition suppression reflects fulfilled perceptual expectations. *Nat. Neurosci.* 11, 1004–1006
43 Damasio, A. *et al.* (2013) Persistence of feelings and sentience after bilateral damage of the insula. *Cereb. Cortex* 23, 833–846
44 Palaniyappan, L. and Liddle, P.F. (2011) Does the salience network play a cardinal role in psychosis? An emerging hypothesis of insular dysfunction. *J. Psychiatry Neurosci.* 36, 100176
45 Blanke, O. (2012) Multisensory brain mechanisms of bodily self-consciousness. *Nat. Rev. Neurosci.* 13, 556–571
46 Botvinick, M. and Cohen, J. (1998) Rubber hands 'feel' touch that eyes see. *Nature* 391, 756
47 Sforza, A. *et al.* (2010) My face in yours: visuo-tactile facial stimulation influences sense of identity. *Soc. Neurosci.* 5, 148–162
48 Lenggenhager, B. *et al.* (2007) Video ergo sum: manipulating bodily self-consciousness. *Science* 317, 1096–1099
49 Ehrsson, H.H. (2007) The experimental induction of out-of-body experiences. *Science* 317, 1048
50 Makin, T.R. *et al.* (2008) On the other hand: dummy hands and peripersonal space. *Behav. Brain Res.* 191, 1–10
51 Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433
52 Ferri, F. *et al.* (2013) The body beyond the body: expectation of a sensory event is enough to induce ownership over a fake hand. *Proc. R. Soc. B Biol. Sci.* 280, 20131140
53 Karnath, H.O. and Baier, B. (2010) Right insula for our sense of limb ownership and self-awareness of actions. *Brain Struct. Funct.* 214, 411–417
54 Schandry, R. (1981) Heart beat perception and emotional experience. *Psychophysiology* 18, 483–488
55 Tsakiris, M. *et al.* (2011) Just a heartbeat away from one's body: interoceptive sensitivity predicts malleability of body-representations. *Proc. R. Soc. B Biol. Sci.* 278, 2470–2476
56 Tajadura-Jimenez, A. *et al.* (2012) The person in the mirror: using the enfacement illusion to investigate the experiential structure of self-identification. *Conscious. Cogn.* 21, 1725–1738
57 Tajadura-Jimenez, A. and Tsakiris, M. (2013) Balancing the 'Inner' and the 'Outer' Self: Interoceptive Sensitivity Modulates Self-Other Boundaries. *J. Exp. Psychol. Gen.* http://dx.doi.org/10.1037/a0033171
58 Moseley, G.L. *et al.* (2008) Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13169–13173
59 Barnsley, N. *et al.* (2011) The rubber hand illusion increases histamine reactivity in the real arm. *Curr. Biol.* 21, R945–R946
60 Kammers, M.P. *et al.* (2011) Feeling numb: temperature, but not thermal pain, modulates feeling of body ownership. *Neuropsychologia* 49, 1316–1321
61 Armel, K.C. and Ramachandran, V.S. (2003) Projecting sensations to external objects: evidence from skin conductance response. *Proc. R. Soc. B Biol. Sci.* 270, 1499–1506
62 Rohde, M. *et al.* (2011) The Rubber Hand Illusion: feeling of ownership and proprioceptive drift do not go hand in hand. *PLoS ONE* 6, e21659
63 Garfinkel, S.N. *et al.* (2013) What the heart forgets: cardiac timing influences memory for words and is modulated by metacognition and interoceptive sensitivity. *Psychophysiology* 50, 505–512
64 Blanke, O. and Metzinger, T. (2009) Full-body illusions and minimal phenomenal selfhood. *Trends Cogn. Sci.* 13, 7–13
65 Friston, K.J. (2005) A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 360, 815–836
66 Dayan, P. *et al.* (1995) The Helmholtz machine. *Neural Comput.* 7, 889–904
67 Hinton, G.E. and Dayan, P. (1996) Varieties of Helmholtz Machine. *Neural Netw.* 9, 1385–1403
68 Clark, A. (2012) Dreaming the whole cat: generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind* 121, 753–771
69 Cechetto, D.F. and Saper, C.B. (1987) Evidence for a viscerotopic sensory representation in the cortex and thalamus in the rat. *J. Comp. Neurol.* 262, 27–45

70 Medford, N. and Critchley, H.D. (2010) Conjoint activity of anterior insular and anterior cingulate cortex: awareness and response. *Brain Struct. Funct.* 214, 535–549
71 Menon, V. and Uddin, L.Q. (2010) Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667
72 Allman, J.M. *et al.* (2010) The von Economo neurons in frontoinsular and anterior cingulate cortex in great apes and humans. *Brain Struct. Funct.* 214, 495–517
73 Evrard, H.C. *et al.* (2012) Von economo neurons in the anterior insula of the macaque monkey. *Neuron* 74, 482–489
74 Khalsa, S.S. *et al.* (2009) The pathways of interoceptive awareness. *Nat. Neurosci.* 12, 1494–1496
75 Craig, A.D. (2011) Significance of the insula for the evolution of human awareness of feelings from the body. *Ann. N. Y. Acad. Sci.* 1225, 72–82
76 Harrison, N.A. *et al.* (2010) The embodiment of emotional feelings in the brain. *J. Neurosci.* 30, 12878–12884
77 Frith, C. (2011) Explaining delusions of control: the comparator model 20 years on. *Conscious. Cogn.* 21, 52–54
78 Fletcher, P.C. and Frith, C.D. (2009) Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* 10, 48–58
79 Adams, R.A. *et al.* (2013) The computational anatomy of psychosis. *Front. Psychiatry* 4, 47
80 Synofzik, M. *et al.* (2010) Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain* 133, 262–271
81 Silani, G. *et al.* (2008) Levels of emotional awareness and autism: an fMRI study. *Soc. Neurosci.* 3, 97–112
82 Seeley, W.W. (2010) Anterior insula degeneration in frontotemporal dementia. *Brain Struct. Funct.* 214, 465–475
83 Kim, E.J. *et al.* (2012) Selective frontoinsular von Economo neuron and fork cell loss in early behavioral variant frontotemporal dementia. *Cereb. Cortex* 22, 251–259
84 Allman, J.M. *et al.* (2011) The von Economo neurons in the frontoinsular and anterior cingulate cortex. *Ann. N. Y. Acad. Sci.* 1225, 59–71
85 Sierra, M. and David, A.S. (2011) Depersonalization: a selective impairment of self-awareness. *Conscious. Cogn.* 20, 99–108
86 Phillips, M.L. *et al.* (2001) Depersonalization disorder: thinking without feeling. *Psychiatry Res.* 108, 145–160
87 Young, A.W. and Leafhead, K.M. (1996) Betwixt life and death: case studies of the Cotard delusion. In *Method in Madness* (Halligan, P.W. and Marshall, J.C., eds), Psychology Press
88 Keller, G.B. *et al.* (2012) Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* 74, 809–815
89 Friston, K.J. (2013) The fantastic organ. *Brain* 136, 1328–1332
90 Friston, K. (2012) Prediction, perception and agency. *Int. J. Psychophysiol.* 83, 248–252