Being a beast machine: the somatic basis of selfhood

Article (Accepted Version)

Being a beast machine: the somatic basis of selfhood

Anil K Seth* and Manos Tsakiris

Anil Seth:
Sackler Centre for Consciousness Science
School of Engineering and Informatics
University of Sussex
BN1 9QJ, UK
+44 1273 678549

Manos Tsakiris:
Lab of Action & Body,
Department of Psychology,
Royal Holloway, University of London &
The Warburg Institute,
School of Advanced Study,
University of London

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

Keywords: interoception, allostasis, predictive processing, embodied selfhood, phenomenology, free energy principle

Abstract

Modern psychology has long focused on the body as the basis of the self. Recently, predictive processing accounts of interoception (perception of the body 'from within') have become influential in accounting for experiences of body ownership and emotion. Here, we describe embodied selfhood in terms of 'instrumental interoceptive inference', which emphasises allostatic regulation and physiological integrity. We apply this approach to the distinctive phenomenology of embodied selfhood, accounting for its non-object-like character and subjective stability over time. Our perspective has implications for the development of selfhood, and illuminates longstanding debates about relations between life and mind, implying – contrary to Descartes – that experiences of embodied selfhood arise because of, and not in spite of, our nature as 'beast machines'.

Being somebody

What does it mean to be a 'self'? While some have argued that there may be no 'thing' that is a self [1], experiences of selfhood are among the most pervasive aspects of human consciousness. Perceptions of the external world come and go, but it is their relation to the experience of 'being an experiencing subject' that gives these perceptions meaning, value, and emotional relevance. How perceptual experiences of 'being a self' are constructed is therefore a key question for cognitive science.

Experiences of having a body, and of being a body, are among the most basic aspects of conscious selfhood [2, 3] upon which higher-level properties of selfhood, such as the experience of being a distinctive individual across time, may rest. Powerful examples of the constructed nature of selfhood are found in experimental and clinical alterations of experiences of body ownership. Experimental manipulations, such as the rubber-hand illusion (RHI) [4] and neuropsychiatric disorders such as asomatognosia or somatoparaphrenia [5, 6] demonstrate that experiences of body ownership do not follow from the mere presence of a physical body-part. Instead, the brain is using available sensory evidence to construct experiences of (dis)ownership that go well beyond the presence or absence of physical body-parts. The RHI shows that a physical body-part is not necessary for a corresponding experience of ownership. In the RHI, such experiences extend to encompass non-self-objects (e.g., fake hands) given appropriate multisensory correlations (e.g., seeing and feeling touch on the fake hand) and sensory inputs that align sufficiently with prior beliefs (e.g., a fake hand that looks sufficiently like a real hand and is roughly where a hand should be). Asomatognosia shows that a physical body-part is not sufficient for experiences of body ownership. Aberrant processing of afferent sensory signals from the affected limbs, or dysfunction of high-level body representations [6], leads to the experience that a (physically present limb) is not part of one's body.

These manipulations of experiences of body ownership encourage interpretation in terms of multisensory integration and predictive self-modelling [7-10]. In this view, such experiences are based on inferences to the best explanation – Bayesian 'best guesses' – which are continually formed and reformed on the basis of neurally-encoded prior expectations and afferent sensory data. The computational principles enabling these inferences are just the same as described for perception of the external world in frameworks like **predictive**

processing [11-13], in which perceptual content is generally assumed to be conveyed by top-down predictions about the hidden causes of sensory signals.

Here, we extend this approach to the embodied self by emphasising two aspects of the relevant predictive models: **interoception** (the sense of the body 'from within', [14]), and the use of predictive *models* for control (**instrumental inference**) rather than for discovery (**epistemic inference**) [15, 16]. We start by reviewing the basics of predictive processing and how it may apply to interoception, especially for purposes of homeostatic regulation [8, 16-18]. We then develop this view in the context of mid-twentieth Century cybernetics and the more recent **free-energy principle**, both of which locate the origin of model-based perception in control and regulation. Our core contribution is to use this perspective to account for the distinctive phenomenology of embodied selfhood, with a focus on its relation to 'objecthood' and the subjective stability of the self across time. We conclude by discussing implications for developmental trajectories [19, 20], psychiatric conditions and for the relationships between consciousness, mind, and life [21-23].

Predictive processing and interoceptive inference

Taking the body as the basis of selfhood highlights the importance of interoceptive sensory channels, which convey information about the global physiological condition of the body [14, 24]. Recently, interoception has been conceptualised within the framework of predictive processing [8, 16, 17, 25, 26]. Just like predictive processing models of vision [27], models of interoceptive inference propose that interoceptive experiences result from probabilistic inference about the causes of viscerosensory inputs, according to Bayesian principles [8, 16, 18]. These models initially focussed on emotional experiences as the relevant consequences of interoceptive inference. On this view, emotional feeling states are shaped by the brain's 'best guess' of the causes of interoceptive signals (Box 1).

The neurocognitive mechanisms of interoceptive inference are assumed to follow the same principles of predictive processing in other modalities. Essentially, the brain embodies and deploys a **generative model** which encodes prior beliefs – in the form of probability distributions – about sensory inputs and their causes in the body and in the world [28]. In popular implementations like **predictive coding** [27], neuronal representations in higher or deeper levels of neuronal hierarchies generate predictions about representations in lower

levels. These descending predictions are compared with lower-level representations to form a prediction error, which is passed back up the hierarchy to update higher-level representations. The exchange of signals between adjacent levels resolves prediction error at every level, resulting in a hierarchically deep neurally-encoded explanation for sensory inputs, and it is this explanation which constitutes the resulting percept [29, 30].

Two aspects of this process are particularly important for what follows. First, sensory prediction error signals are precision-weighted such that signals with high (expected) precision (inverse variance) have greater influence in updating descending predictions. Note that the precision of sensory data, just like its mean (or any other distributional property), has to be inferred. Inferred precision depends both on the empirical variance of the sensory data and on prior expectations about precision. The optimisation of precision-weighting, through changes in precision-related priors (precision expectations) is frequently associated with attention [31]. Intuitively, paying attention to sensory data is equivalent to increasing its expected precision so that the sensory data has greater impact on perception.

Equally important is that sensory prediction errors can be minimized by performing actions to change sensory data, as well as by updating predictions. Minimizing prediction error through action is called **active inference** [32]. For example, visual prediction error can be reduced by moving one's eyes until a prediction is fulfilled: if I expect to see a nose but am currently perceiving an ear, a simple saccade will often make the nose prediction come true. Actions themselves can be thought of as the fulfilment of proprioceptive (or oculomotor) predictions [32, 33]: an intended movement occurs by predicting its proprioceptive consequences. Crucially, this applies also to interoception, where interoceptive prediction errors can be minimized through autonomic reflexes, or, more broadly, 'intero-actions' [34]. In general, active inference depends on the ability of generative models to make predictions about the sensory consequences of specific actions [35].

Putting these features together, one can see that active inference provides a means for control or regulation of inferred causes (the hidden or latent variables from which sensory signals originate). Given some sensory prediction error, whether predictions are updated, or whether actions are performed to change the sensory data, depends on precision weighting. Decreasing the expected precision of sensory prediction errors will lead to predictions dominating, so that prediction errors will be resolved through action. For example, motor

actions are elicited when descending proprioceptive predictions set equilibrium points which engage classical reflex arcs [33, 35]. This entails lowering the expected precision of proprioceptive prediction errors, corresponding to diminished attention to proprioceptive and kinaesthetic sensations. This in turn explains observations of increased sensory thresholds in these modalities during movements [36].

This is a brief description of predictive processing and active inference. We next turn to its relevance for embodied selfhood by revisiting some seminal concepts in cybernetics, and some recent developments in theoretical neurobiology, which together emphasise the importance of predictive modelling for control.

From essential variables to instrumental (control-oriented) inference

All living organisms attempt to maintain their physiological integrity in the face of danger and opportunity. Arguably, this is the basic evolutionary and functional imperative for having a brain. In the 1950s the cybernetician W. Ross Ashby formalized this idea in terms of second-order homeostasis of **essential variables**. In physiological settings, these variables correspond to quantities like blood pressure, heart rate, blood sugar levels and the like: quantities which must remain within tight bounds in order for an organism to survive. In Ashby's framework, when essential variables move outside organism-specific viability limits (following a breakdown in first-order homeostatic processes, like simple feedback), adaptive processes are triggered which re-parameterize the system until it reaches a new equilibrium in which homeostasis is restored [37]. Ashby called this (second-order) process 'ultrastability': an ultrastable system is capable of finding a new stable configuration with its environment, even given perturbations sufficient to disrupt ongoing homeostatic processes. In early descriptions of ultrastability, second-order re-parameterization was implemented as a random process. For example, Ashby's famous-at-the-time 'homeostat' would randomly explore different settings once an essential variable had transgressed its bounds, and would continue to do so until first-order homeostasis had been restored. However, random exploration of parameter settings is inefficient and biologically implausible. (The mathematician Norbert Wiener called the homeostat 'one of the great philosophical contributions of the present day'; see http://blogs.bl.uk/science/2016/04/the-thinking-machine.html).

A more useful solution is provided by models capable of explicitly inferring bodily states (essential variables) and their homeostatically-relevant trajectories over time, and of acting on these states in order to assure ongoing physiological integrity [37]. In embodied settings, model-based control is mandated by several factors. First, the hidden causes of interoceptive signals – i.e. the targets of physiological regulation, the values of essential variables – are not directly available to the brain's control mechanisms and must be inferred. Indeed, this is the primary rationale for proposals of interoceptive inference [8, 16, 17]. Second, model-based control allows anticipatory responses, since models allow inferences about future bodily states and their trajectories, and support conditional predictions about these states given specific (autonomic or motor) actions. In physiological settings, anticipatory control can be critical. Waiting for tightly regulated quantities like blood acidity to exceed their bounds before engaging compensatory responses may be lethal [38]. Third, and related, hierarchical models allow anticipatory control to play out across multiple levels, so that regulation at one level may be temporarily relaxed (or altered through imposing a new set-point) in order for homeostasis to be preserved at higher levels or over longer timescales. For example, transient changes in blood pressure regulation may be necessary to enable fight-flight responses when encountering a predator [17, 37, 38]. Simply standing up from your desk imposes similar anticipatory demands, though less dramatically. Altogether, these capabilities describe a transition from (first-order) homeostasis to the more general process of **allostasis**: the regulation of bodily states through change [37, 39, 40].

This notion of allostatic regulation captures the core idea of instrumental (control-oriented) interoceptive inference. In just the same way as described for motor actions, instrumental interoceptive (active) inference involves descending interoceptive predictions being transcribed into physiological homeostasis by engaging autonomic reflex arcs (intero-actions). However, whereas motor actions may serve different goals over time, requiring ever-changing changing proprioceptive set-points, physiological homeostasis entails maintaining physiological essential variables within tight ranges of viability at all times. We will see later that this difference has implications for the subjective stability of embodied selfhood.

The general idea that the brain encodes models for predictive inference and allostatic regulation is not new. In 1970, Ashby, with Roger Conant, proposed the influential "**good regulator theorem**" which states that "every good regulator of a system must be a model of

that system" [41]. The overlooked **perceptual control theory** of Powers similarly emphasises control, arguing that the purpose of behaviour is to regulate perceptual variables: "control systems control what they sense, not what they do" [42]. For example, when catching a cricket ball, experienced cricketers move so as to control a perceptual variable (the rate of change of the tangent of the angle of elevation to the ball), facilitating their arrival in the right place at the right time [43].

More recently, allostatic regulation has become central to accounts of perception and action under the **free-energy principle**, which starts from the premise that all systems which preserve their identity over time must resist a tendency towards dispersion of their internal states, by minimizing the long-run unexpectedness or entropy of these states – which under some simplifying assumptions corresponds to minimization of precision-weighted sensory prediction errors [44] (Box 2). Explicit links between allostatic regulation and interoceptive inference have also now been elaborated, highlighting implications for functional neuroanatomy [45, 46], and psychiatric disorders like depression [18, 37, 47]; see Box 3.

These various perspectives all highlight a subtle but significant distinction between a system 'being a model', in the sense that it can be described in a model-based way, and 'having a model', in the sense of explicitly encoding a probabilistic model (of the hidden causes of sensory signals, their trajectories, and conditional dependencies on actions). The potential relevance of this distinction for the phenomenology of selfhood is discussed further in Box 4.

The argument so far is as follows. Basic imperatives towards sustained physiological integrity mandate the implementation of control-oriented predictive models. These hierarchically organised models implement active inference on interoceptive signals to enable allostatic regulation. Lower hierarchical levels implement autonomic reflexes, while higher levels recruit multimodal and amodal prior expectations about the physiological consequences of actions, supporting behavioural regulation of physiological states over longer periods of time and in different contexts. Just as visual experience can be understood as the content of visual predictions, interoceptive experience can be thought of as the content of the joint set of predictions geared towards allostasis. With these pieces in place, we can now examine how this perspective on interoceptive inference sheds new light on the phenomenology of embodied selfhood.

The phenomenology of 'being a body'

When considering the phenomenology of selfhood, it is not enough to say that emotional and self-related experiences are the way they are (and are different to, for example, visual experiences) because they emphasise predictions about interoceptive (rather than visual) signals. Instead, it is helpful to consider the nature of predictions associated with interoceptive inference, especially their control-oriented (instrumental) bias.

Epistemic and instrumental inference

On the view we propose, the function of perception is not to recover a veridical, action-independent representation of the external environment or body. Instead, predictive perception – in any modality – is ultimately geared towards driving actions that preserve physiological integrity of the organism. In other words, we do not perceive the world (and self) as it is, but as it is useful to do so. This may involve systematically 'misperceiving' the environment, by criteria of veridicality. At one level, the ability to elicit motor or autonomic actions through the fulfilment of descending predictions could not happen without suppressing ('veridical') perception of the current state-of-affairs, through sensory attenuation. (We recognise that empirical evidence for interoceptive sensory attenuation remains to be found, and that this process may be realised differently than in the motor domain due to different time constants [36].)

At another level, one can distinguish between 'epistemic' and 'instrumental' varieties of active inference. **Epistemic inference** prescribes 'information-gathering' actions that enhance the predictive capabilities of a model, in order to enhance its regulatory capabilities in the long run [48, 49]. **Instrumental inference** (equivalently, control-oriented inference) prescribes actions that exploit these models for ongoing control and regulation [15, 50, 51]. Both concepts invite description as 'inference' since both involve 'going beyond the data' in virtue of generative modelling [13]. They are also both forms of active inference since they both prescribe actions. Note that there is no necessity that predictive models most useful for control are those that represent sensory signals and their causes in an action-independent, veridical manner [15, 52, 53]. This motivates a further distinction, or spectrum, between epistemic actions that enhance a predictive model in terms of veridicality, and those deployed to improve its regulatory capabilities. Put this way, the deep interdependencies between

perception and action can be understood in terms of an ever-shifting balance between the discovery of behaviourally (allostatically) relevant features of the environment (epistemic inference) and the control or regulation of the causes of sensory signals (instrumental or control-oriented inference).

The embodied self is not just an object

The distinction between epistemic and instrumental inference helps explain aspects of the phenomenology of embodied selfhood, especially in terms of its relation to 'objecthood'. Consider that in visual experience, perceptual scenes seem organised, to a large extent, into discrete objects and the spaces between them. In some sense, the experience of an object includes the perception of surfaces that are not directly represented in sensory data. When we perceive an object, we perceive it as having an external existence, with a 'back and sides' [54], as 'really existing' out there in the world [55, 56].

Through some modalities (e.g., vision) the body too can be experienced as an object. You may (visually) perceive the hand in front of me as *your hand* – an experience of body ownership which can extend to the entire body-as-object [57, 58], and which can be influenced by interoceptive signals [59, 60]. However, many aspects of self-related phenomenology are not like this. William James put it this way: "contrary to the perception of an object, which can be perceived from different perspectives or even cease to be perceived, we experience 'the feeling of the same old body always there'" [61] (p.242). For example, affective experiences - like emotions - do not occupy a volume in space, nor do they have a back-and-sides.

Basic experiences of 'embodied selfhood' are even more challenging to describe [2]. There is a very low-level aspect of embodied self-experience perhaps best described as the experience of being a living organism, as opposed to 'owning' a particular body, which seems to resist easy analysis in terms of the sort of perceptual inference responsible for exteroceptive perceptual scenes. This inchoate (and 'transparent', see [62]) phenomenology of 'being a body' describes a background experience of selfhood that shades into mood and emotion at one end, and into experiences of body ownership at the other. This deeply rooted aspect of experienced embodiment involves no strong component of objecthood. Crudely put, we do not experience 'being a body' in terms of the spatial arrangement of our internal organs as objects.

This difference in phenomenology can be explained by the following hypothesis, based on the distinction between epistemic and instrumental inference. Your visual experience of a mug has the phenomenological character of objecthood because your brain is making epistemic predictions about how mug-related sensory signals would change given this-or-that (epistemic) action. You perceive a mug as having a back because your brain encodes predictions that the back would come into view, if you rotated the mug. More generally, perceptual inferences have the phenomenological character of objecthood when the underlying generative models deploy a rich set of epistemic predictions about sensory signals you would encounter were you to make (conditional) or if you had made (counterfactual) this-or-that action [55, 56]. This is simply a predictive processing version of the fundaments of sensorimotor contingency theory [63] which argues that the phenomenology of objecthood comes from the 'mastery' of the relevant sensorimotor contingencies.

In interoceptive inference, however, actions - whether autonomic or motoric - serve predominantly to regulate interoceptive sensations [18, 47] (c.f. perceptual control theory [42]). Perceptual predictions relevant to interoceptive inference have more to do with instrumental predictions about the physiological consequences of actions, than about (epistemically) discovering more about some particular external or internal state-of-affairs. Interoceptive inference therefore marks a different balance from typical exteroceptive inference on the trade-off between refining a model (epistemic inference) and deploying a model for regulation (instrumental inference). Note that both epistemic and instrumental inference involve conditional or counterfactual predictions, but in different ways: in the former case, to predict how an action would improve the model; in the latter, to predict how an action would affect physiological homeostasis, given a model.

Based on these distinctions, we propose that instrumental inference undergirds a different phenomenology than epistemic inference related to discovery. Instead of delivering a phenomenology of objecthood, instrumental (control-oriented) interoceptive inference plausibly underlies a phenomenology related to the evaluation of the allostatic consequences of regulatory actions. A non-localised, non-object-based phenomenology associated with both mood and emotion, and with the pre-reflective (i.e., non-reflexive) self-related experience of being an embodied organism [15, 18].

The stability of the self over time

Another striking aspect of embodied selfhood, at least in non-pathological situations, is its subjective stability over time. Perceptions of the world come-and-go, but experiences of selfhood seem stable and continuous over many different time scales. How can this be accounted for?

One possibility is that the hidden causes of self-related perceptions may indeed *be* more stable than the hidden causes of world-related signals, simply due to their origin in a relatively unchanging, and allostatically controlled, milieu. As mentioned previously and in contrast to active inference in the motor system, instrumental interoceptive inference requires maintaining physiological essential variables within tight ranges of viability across time. This entails precise prior expectations that these variables and their trajectories remain within such stable ranges, with corresponding sensory attenuation of interoceptive signals. The resulting interoceptive perceptions will therefore be drawn towards stable inferences about self-related variables and their trajectories.

In addition, as phenomena like change blindness amply demonstrate [63], perception of change is not the same as change of perception. The subjective stability of selfhood, at some levels, may reflect an adaptive form of 'self-change-blindness'. In this view, allostatic regulation, and apparent goal-directed behaviour ultimately motivated by such regulation, may depend on *not* perceiving that the relevant aspects of selfhood (i.e., the targets for allostatic regulation) are changing, in order to provide stable targets for instrumental inference. This would be the case even if these aspects of self, and perceptions of self, are in fact changing. This is an application of the idea, outlined above, that instrumental inference may require systematic 'mis-perception' of the hidden causes of sensory signals. Put simply, we will be better able to maintain our physiological and psychological integrity and identity over time if we do not (expect to) perceive ourselves as continually changing. This applies at many levels of selfhood, from tightly regulated aspects of the interoceptive self, to the preservation of a stable personal identity during temporally extended social interactions over days, months, and years – including interactions with oneself via recall of episodic and autobiographical memories and planning for the future. Speculatively, breakdowns in such self-change-blindness may be implicated in psychiatric disorders in which the stability of the

self becomes subjectively unreliable, such as in schizophrenia, dementia and delirium, or multiple personality disorder.

Concluding remarks and future perspectives

Experiences of selfhood range from basic experiences of 'being' and 'having' a body, up to reflective self-awareness and the social self [8, 64] (see Box 4). We have proposed that these experiences are grounded in processes of instrumental (control-oriented) interoceptive inference that underpin allostatic regulation of physiological essential variables. This perspective draws together perceptual inference schemes, such as predictive processing and active inference, with sensorimotor theory [42, 65] and concepts from mid-twentieth-century cybernetics [41, 66] that emphasise model-based control, with deep links to allostasis and physiological integrity finding formal expression in the free energy principle ([44, 67], Box 2). This perspective accounts for distinctive aspects of the phenomenology of selfhood, including its relationship to 'objecthood' [55] and its subjective stability over time. It may also shed light on aspects of aberrant self-experience (Box 3), and its developmental trajectory – especially in relation to caregiver dynamics (Box 5).

The story has involved two primary 'inversions' with respect to the classical view of perceptual content arising from 'bottom-up' (or outside-in) elaboration of sensory signals [68, 69]. First, perceptual content is conveyed by top-down (or inside-out) predictions about the causes of sensory signals, rather than by the sensory signals themselves. This is common to all most if not all predictive processing frameworks [11, 12]. (Note that we avoid the terms 'feedforward' and 'feedback' since these carry implications about error signals which do not match the architecture of predictive processing [70].) Second, interoceptive inference, and instrumental inference more broadly, should not be considered as a generalisation of predictive coding from exteroceptive modalities like vision. Instead, perceptual content in all modalities, including modalities such as vision, is a consequence or generalisation of a fundamental imperative towards physiological regulation [15]. Seen this way, all perceptual content is underpinned by inferential mechanisms that have a functional, ontological, and phylogenetic basis in allostasis.

The deep physiological roots of instrumental inference gesture towards a third 'inversion', which has to do with the debated connection between 'life' and 'mind' [71], and which traces back to Enlightenment discourse about the nature of the soul and the relevance of the body

[72]. For Descartes, non-human animals were 'beast machines' without souls or conscious experiences of any kind, at least without any kind of experience warranting moral status. Their flesh-and-blood nature was highlighted as irrelevant to the presence of consciousness or 'soul'. Writing after Descartes, in 1748, Julien de La Mettrie took this idea to its extreme: if animals are beast machines then so are humans, since humans are also animals ([73]; see also [72, 74]). Our view suggests the opposite: that there are intimate connections between the functional imperatives imposed by our physiological reality – by 'the drive to stay alive' that animates all living creatures – and the predictive machinery that implements instrumental interoceptive inference.

This view pushes back against popular views of mind and self as substrate-independent forms of information processing [69, 75]. At minimum, it suggests that mind and self cannot be understood without deep appreciation of the constraints and opportunities afforded by embodiment and allostasis. More radically, it underpins a strong continuity between life, mind, and consciousness [21-23, 76]. The implications of this line of thinking raise many questions in areas from artificial intelligence to computational psychiatry (see **Outstanding Questions**). But the basic message is simple. We perceive the world around us, and ourselves within it because of, and not in spite of, the fact that we are beast machines.

Acknowledgements. AKS is grateful for support to the Dr. Mortimer and Theresa Sackler Foundation, which supports the work of the Sackler Centre for Consciousness Science, and to the Canadian Institute for Advanced Research (CIFAR) Azrieli Programme on Brain, Mind, and Consciousness. MT is supported by the European Research Council Consolidator Grant (ERC-2016-CoG-724537) under the FP7 for the INtheSELF project and by the NOMIS Foundation Distinguished Scientist Award. The authors are grateful to Hugo Critchley, Klaas Enno Stephan, Alec Tschantz, and our anonymous reviewers for helpful comments.

Box 1: Interoceptive inference, emotion, and mood

Existing models of interoceptive inference primarily target emotional and affective aspects of perception and self [8, 17, 25, 47]. These models propose that interoceptive experience results from inference on the hidden causes of interoceptive signals, by analogy with predictive processing models of exteroceptive perception [11, 12] and echoing the early ideas of von Helmholtz of perception as 'unconscious inference' [29, 30]. They extend earlier theories of emotion in two important ways. First, like 'appraisal' theories [77] they emphasize the importance of context in emotional processing, but with the advantage of eliminating any bright line separating emotion (or perception) from cognition. Instead, interoceptive experience is determined by inferential processes operating across multiple hierarchical levels and encompassing multiple modalities. Second, instead of associating each emotion with a discrete neuronal circuit [78], they view emotions as *constructed* by neural processes, such as perceptual inference and memory, which reflect principles of structural and functional organisation that generalise beyond emotion itself [79].

As with the relationship between exteroceptive (e.g., visual) perceptual content and predictive processing, it remains an open question as to how interoceptive experiences map onto the computational machinery of interoceptive inference (see e.g., [80, 81] for examples in vision and audition). Interoceptive experience might reflect the posterior belief directly [8], the trajectories of interoceptive prediction error (or free energy, see Box 2) over time [82], the precision (certainty) of the predicted somatic consequences of (motoric or autonomic) actions [18, 82], and/or hyper-priors over this precision. Clark and colleagues recently suggested a hierarchical arrangement in which emotional experience reflects precision, while mood depends on the hyper-prior over this precision (i.e., expected precision) [83]. In this interesting view, short term fluctuations in precision (emotional responses) are constrained by mood-related hyperpriors that encode their long-term average. Beyond this, we note that direct empirical evidence for interoceptive inference (e.g., interoceptive prediction errors) remains scarce ([84], see Outstanding Questions).

Which aspects of (active) interoceptive inference shape *conscious* emotional experience? While this question also remains open, it is tempting to speculate that expectations at higher (deeper) levels of neuronal hierarchies are more likely to be implicated, largely because their predictions are domain general and can therefore be articulated through autonomic or motor reflexes.

Box 2: The free energy principle

The free energy principle (FEP), as described principally by Karl Friston, is the most ambitious of theoretical frameworks related to predictive processing [28, 44]. According to the FEP all organisms, simply by virtue of existing, are mandated to minimize the entropy, dispersion, or 'atypicality' of their states. In other words, organisms inhabit states in which they expect to be in – where 'expect' is interpreted in terms of neuronally-encoded probability distributions, not personal-level psychological beliefs. This basic condition on the nature of living organisms stems from their need to resist the tendency towards disorder imposed by the second law of thermodynamics, and is taken to apply to all features of living systems - from their gross morphology to fine details of cortical microcircuitry – as well as at timescales from the neuronal to the phylogenetic [23]. Entropy is the long-term average of (information-theoretic) surprise, which cannot be directly measured. The FEP therefore supposes that organisms minimize a proxy or upper-bound, which is called the (variational) 'free energy'. Under simplifying assumptions (including Gaussian distributions and independence/factorisation of time scales), free energy is equivalent to precision-weighted prediction error, which means that schemes like predictive coding and active inference [18, 32] become process theories (implementations) under the FEP [44].

While the mathematical details of the FEP are complicated (see [67] for a recent review), the basic message is simple. It is that the computational machinery of predictive perception – and more importantly control-oriented predictive regulation (instrumental active inference) – stems from basic physical principles that apply to all living systems (perhaps even to all systems that can be said to exist, or to persist; the notion of a **Markov blanket** becomes relevant here [23, 85]), which entail that such systems must maintain themselves within a limited repertoire of states.

Although derived from different traditions, the FEP shows clear parallels with cybernetic theories which emphasise feedback, control, and predictive modelling – in particular the 'good regulator theorem' [41]. Indeed, process theories which specify explicit generative models for active inference usefully address the distinction between 'being a model' and 'having a model' which is left ambiguous under this earlier theorem (Box 4). Such process theories also provide a recipe for agent-based models [48, 86], some of which are illustrating the conditions under which generative models that implement effective predictive regulation will depart from 'veridical' models of the hidden causes of sensory signals [53, 67]. Future developments

of these models will shed further light on the distinction between epistemic and instrumental inference [15].

Box 3: Psychopathology of disrupted interoceptive inference

Interoceptive inference provides new opportunities to relate symptom expression to altered neurocomputational mechanisms, in a range of psychiatric and psychological conditions [34, 87]. Early proposals interpreted anxiety as a consequence of chronically elevated interoceptive prediction error [88], and associated dissociative conditions like depersonalisation/derealisation [89] with imprecise interoceptive predictions [25]. While these proposals remain worth investigating, a control-oriented perspective suggests additional targets.

Quattrocki and Friston have suggested that features of autistic spectrum disorder (ASD) arise from developmental abnormalities in the modulation of interoceptive prediction errors [90]. In their view, aberrant modulation of the (expected) precision of interoceptive prediction errors during interactions between infants and caregivers prevent the infant from developing the hierarchically-deep generative models able to properly attribute hidden causes of interoceptive signals to 'self' and to 'other' (Box 5). In a related non-developmental view, Palmer and colleagues suggested that social symptoms in ASD reflect a diminished set of conditional or counterfactual predictions relating to inferred states-of-mind of others [91] (see also [92]). Such inferentially-impoverished mentalising may lead to diminished 'perceptual presence' of other minds, just as perceptual presence in vision may be diminished when generative models cannot support rich predictions about the sensory consequences of actions [55] (see Section: The body is not just an object). These ideas provides interesting contrasts to accounts based on 'theory of mind' [93], and may explain the autonomic hypersensitivity and other interoceptive symptoms that are often observed in individuals with autism – as well as difficulties engaging with exteroceptive prosocial cues [94, 95].

Another important application is to depression. Barrett and colleagues consider depression an allostatic disorder in which the brain becomes pathologically insensitive to prediction error signals and, consequently, less effective in terms of (metabolic) energy regulation [47]. They argue that a metabolically inefficient internal model of the 'body in the world' accounts for a wide variety of symptoms and aetiologies associated with depression, including its pervasive negative affect and association with apathy and fatigue. In a related view, Stephan and colleagues propose that fatigue and depression are sequential responses to interoceptive experiences of dyshomeostasis (fatigue) and subsequent metacognitive beliefs about low

allostatic self-efficacy (depression) [37], extending early ideas based on 'learned helplessness' and generalised loss-of-control [96]. Clark and colleagues similarly emphasize precision, associating major depression with being certain about encountering uncertain environments, thereby precluding effective allostatic regulation [83]. These proposals, while distinct, share features – and all emphasise interoceptive inference. Future work employing computational psychiatry methods could arbitrate among them, as well as isolating opportunities for patient stratification and intervention [97], and extension to radical disturbances of selfhood such as Cotard's syndrome, where people believe that they do not exist [25, 98].

Box 4: Being a model versus having a model

Is there any substantive difference – when it comes to experiences of embodiment and selfhood – between systems that explicitly deploy a predictive model, as compared to systems that are merely aptly described in model-based terms? Most process theories (implementations, e.g., predictive coding) relevant to interoceptive inference imply that neural states encode the parameters (sufficient statistics) of a generative model which maps between sensory data and their hidden causes [67]. However, broader theoretical frameworks like the free energy principle (Box 2) and the good regulator theorem [41], which emphasize regulation, do not in themselves mandate the existence of an explicitly encoded generative model; only that systems behave in ways that are well described this way [85]. Since regulation can in principle occur with or without explicitly encoded generative models, the *way in which* systems engage in allostatic regulation may have consequences for self-related phenomenology.

Simple forms of regulation such as first-order homeostatic feedback can be thought of as simply 'being a model'. More complex regulation, involving inferential, anticipatory (forecasting), and flexible control, may require explicit generative modelling. This can be described as 'having a model'. This distinction recalls an old debate in cognitive science as to whether systems explicitly represent properties of their environment, or whether they merely act as if they do [99]. In the context of interoceptive inference, explicit generative modelling is particularly relevant in virtue of supporting predictions about the future somatic effects of actions (forecasting), in the context of the intrinsic dynamics of the body and the environment (which includes other embodied agents, see Box 5) [34].

While this distinction is unlikely to remain sharp, it helps bring into view a range of possible control structures of increasing model-based explicitness and hierarchical depth. We suggest that as control mechanisms develop in these directions, the associated phenomenology of selfhood may develop from experiences of being an embodied organism, to experiences of mood and emotion, pre-reflective experiences of selfhood and 'mineness', and finally to explicit self-awareness, metacognitive insight, reflective self-awareness, and social aspects of selfhood. (See [62] for a discussion in terms of representational 'transparency' and 'opacity'.)

Box 5: Development of the interoceptive self

If interoceptive inference is needed for keeping the organism within regimes of physiological viability, at the beginning of human life this process is critically dependent on caregivers. Human infants are born lacking the ability to perform autonomously the actions needed for addressing their internal sensations and needs. From eating and drinking to thermoregulation and sleep, their bodily regulation depends on others; specifically, on intersubjective carer-infant embodied and affective interactions.

Such intersubjective approaches have recently been extended to interoception [19]. The development of visceral and emotional neural circuitry depends on a caregiver-infant relationship [100, 101], often conceptualized as homeostatic regulation [101]. The first months post-partum are characterized by relative instability of key cardiovascular variables (e.g. heart rate variability, vagal tone) that become moderately stable by the end of the first year [100]. Importantly, their levels depend on caregiving [102] such as parent-infant contingency during interaction [103], and are predictive of self-regulation abilities at the age of three [104].

Beyond homeostasis *per se*, such interactions enable the infant to learn to associate specific homeostatic needs (e.g., pain) and their behavioural expression (e.g., crying) to contingent allostatic responses from the carer (e.g., soothing rather than feeding; see [19]). Given the infant's inability to perform the required allostatic actions, it is the caregiver's task to do so (see Glossary and also [105]). This depends on their ability to correctly infer the hidden causes of the infant's putative interoceptive prediction error and provide an appropriate response. The accumulation of such responses, derived from precise interoceptive predictions performed by the carer on behalf of the infants, will eventually lead to the construction, by the infant, of a predictive model of their interoceptive body. Consequently, imprecise inference of the infant's hidden causes of interoceptive changes may hinder the development of an allostatically adequate model. On this view, early coupled intersubjective embodied iterations provide the necessary precision-weighting, so that it is with others that we develop a sense of ourselves 'from within'.

It has been recently shown that parental interoceptive sensitivity, measured neurally (as reflected in anterior insula activity) as well as behaviorally, during the first months of parenting was predictive of their children's somatic symptoms six years later [106]. These

findings provide tentative support to the crucial role that carer-child interactions play in supporting interoceptive development, and they chart two pathways that may shape interoceptive sensitivity cross-generationally. The first, consisting of the amygdala and oxytocin system, supports attention to arousal modulations in response to social cues. The second, involving anterior insula, supports higher-order interoceptive representations that underpin embodiment and self-awareness.

Such interoceptive approaches to self-development coupled with new methods for assessing interoceptive sensitivity in infants, such as the Infant Heartbeat Task (iBEAT [20]), will enable us to study the ontogenetic development of interoception, mentalization and metacognition of bodily experience [92, 107], and will advance our understanding of developmental disorders such as autism [90, 108], eating disorders [109] and affective disorders [18].

Outstanding questions

- What are the specific neurophysiological signatures of interoceptive predictions and prediction errors, and where are they localized in the brain?
- Can the proposed phenomenological differences between epistemic and instrumental (active) inference be tested in a non-interoceptive domain, like vision?
- Given the importance of interoception and physiological regulation in embodied selfhood, what might be the roles of epigenetics, the immune system, the microbiome, and other somatic systems on the interoceptive predictions underlying experiences of self and affect?
- If perceptions of self and world arise from fundamental imperatives towards allostatic regulation, what are the prospects for artificial intelligence and machine consciousness? Would a robot need to be 'alive' in order to be ascribed, or to ascribe itself, with selfhood?
- Can new cognitive behavioral therapy treatments be grounded in refining people's predictive models of their ability to allostatically regulate, for instance by training metacognitive awareness of interoceptive signals?
- Are the targets of allostatic control set points or ranges? And if the latter, can ideas about so-called 'rein-control' (two complementary control systems, like the reins used to steer horses) be mapped to the neuroanatomy of allostasis?
- How does the ontogenetic development of interoceptive awareness link to other dimensions of self-awareness in early life, such as self-recognition and how does it change during other critical developmental periods such as adolescence?

Glossary

Active inference: An extension of predictive processing, 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 [28, 32, 67].

Allostasis: A form of regulation which emphasizes the process of achieving stability through change, for example by the dynamic and anticipatory allocation of resources to ensure the stability of core regulatory targets. The precise relationship between allostasis and homeostasis is still debated (e.g., [39]).

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 [61].

Essential variables: Physiological quantities which must remain within specific bounds for an organism to remain viable (to stay alive). The term is associated with the 20th Century cybernetician W. Ross Ashby [66].

Epistemic inference: A subset of active inference where actions serve primarily to enhance generative models, through discovering more about the hidden causes of sensory signals, therefore enabling enhanced prediction error minimization in the long run.

Free energy principle: A generalization of predictive processing according to which organisms minimize a lower bound on the entropy of sensory signals (the free energy). Under some assumptions free energy translates to precision-weighted prediction error [44].

Generative model: A probabilistic model linking (hidden) causes and data, usually specified in terms of the likelihoods (of observing some data given their causes), and priors (on these causes). Generative models can be used to generate fictive data samples of the sort needed to guide active inference.

Good regulator theorem: A thesis from cybernetics which claims to show, under broad conditions, that 'every good regulator of a system must be a model of that system' [41].

Interoception: The sense of the internal physiological condition of the body [14]. Interoception, as used here, encompasses afferent sensory signals (interosensations) from the viscera as well as low-level monitoring of blood chemistry and sensations evoked by affective touch or pain, interoceptive perceptions which comprise perceptual inferences about the body states that cause interosensations, and interoactions which implement allostatic control through autonomic reflexes [34].

Instrumental inference: A subset of active inference in which actions serve primarily to regulate perceptual variables (and their hidden causes). Equivalently, control-oriented inference.

Markov blanket: A Markov blanket defines the boundaries of a system in a statistical sense, so that variables within the blanket are conditionally independent of those outside the blanket, and *vice versa* [85].

Perceptual control theory: A relatively overlooked theory, developed by William Powers and based on principles of hierarchically-nested negative feedback, which interprets behavior as implementing the control of perceptual variables – rather than perception controlling behavior [42].

Predictive coding: A data processing strategy whereby signals are represented by generative models. Predictive coding is typically implemented by message-passing architectures in which top-down signals convey predictions and bottom-up signals convey prediction errors [27].

Predictive processing: A theoretical framework in which perception, action and cognition depend on the deployment of multi-level generative models to predict the incoming sensory barrage [11]. Predictive coding is an implementation of predictive processing.

References

1. Metzinger, T., *Being No-One*. 2003, Cambridge, MA: MIT Press.
2. Blanke, O. and T. Metzinger, *Full-body illusions and minimal phenomenal selfhood.* Trends in cognitive sciences, 2009. **13**(1): p. 7-13.
3. Gallagher, I.I., *Philosophical conceptions of the self: implications for cognitive science.* Trends Cogn Sci, 2000. **4**(1): p. 14-21.
4. Botvinick, M. and J. Cohen, *Rubber hands 'feel' touch that eyes see.* Nature, 1998. **391**(6669): p. 756.
5. Vallar, G. and R. Ronchi, *Somatoparaphrenia: a body delusion. A review of the neuropsychological literature.* Exp Brain Res, 2008.
6. Feinberg, T.E., et al., *The neuroanatomy of asomatognosia and somatoparaphrenia.* J Neurol Neurosurg Psychiatry, 2010. **81**(3): p. 276-81.
7. Limanowski, J. and F. Blankenburg, *Minimal self-models and the free energy principle.* Front Hum Neurosci, 2013. **7**: p. 547.
8. Seth, A.K., *Interoceptive inference, emotion, and the embodied self.* Trends Cogn Sci, 2013. **17**(11): p. 565-73.
9. Apps, M.A. and M. Tsakiris, *The free-energy self: A predictive coding account of self-recognition.* Neurosci Biobehav Rev, 2014. **41C**: p. 85-97.
10. Limanowski, J., *What can body ownership illusions tell us about minimal phenomenal selfhood?* Front Hum Neurosci, 2014. **8**: p. 946.
11. Clark, A., *Whatever next? Predictive brains, situated agents, and the future of cognitive science.* Behav Brain Sci, 2013. **36**(3): p. 181-204.
12. Hohwy, J., *The Predictive Mind*. 2013, Oxford: Oxford University Press.
13. de Lange, F.P., M. Heilbron, and P. Kok, *How do expectations shape perception?* Trends Cogn Sci, 2018.
14. Craig, A.D., *Interoception: the sense of the physiological condition of the body.* Curr Opin Neurobiol, 2003. **13**(4): p. 500-5.
15. Seth, A.K., *The cybernetic bayesian brain: from interoceptive inference to sensorimotor contingencies*, in *Open MIND*, J.M. Windt and T. Metzinger, Editors. 2015, MIND Group: Frankfurt A .M. p. 1-24.
16. Barrett, L.F. and W.K. Simmons, *Interoceptive predictions in the brain.* Nat Rev Neurosci, 2015. **16**(7): p. 419-29.
17. Pezzulo, G., F. Rigoli, and K. Friston, *Active Inference, homeostatic regulation and adaptive behavioural control.* Prog Neurobiol, 2015. **134**: p. 17-35.
18. Seth, A.K. and K.J. Friston, *Active interoceptive inference and the emotional brain.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2016. **371**(1708).
19. Fotopoulou, A. and M. Tsakiris, *Mentalizing homeostasis: The social origins of interoceptive inference.* Neuropsychoanalysis, 2017. **19**(1): p. 3-28.
20. Maister, L., T. Tang, and M. Tsakiris, *Neurobehavioral evidence of interoceptive sensitivity in early infancy.* Elife, 2017. **6**.
21. Godfrey-Smith, P.G., *Complexity and the function of mind in nature*. 1996, Cambridge, MA: MIT Press.
22. Thompson, E., *Mind in life: Biology, phenomenology, and the sciences of mind*. 2007, Cambridge, MA: Harvard University Press.
23. Friston, K.J., *Life as we know it.* J R Soc Interface, 2013. **10**(86): p. 20130475.
24. Denton, D., *The primordial emotions: The dawning of consciousness*. 2006, Oxford: Oxford University Press.
25. Seth, A.K., K. Suzuki, and H.D. Critchley, *An interoceptive predictive coding model of conscious presence.* Frontiers in Psychology, 2011. **2**: p. 395.

26. Marshall, A.C., A. Gentsch, and S. Schutz-Bosbach, *The Interaction between Interoceptive and Action States within a Framework of Predictive Coding.* Front Psychol, 2018. **9**: p. 180.
27. Rao, R.P. and D.H. Ballard, *Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects.* Nat Neurosci, 1999. **2**(1): p. 79-87.
28. Friston, K.J., *The free-energy principle: a rough guide to the brain?* Trends Cogn Sci, 2009. **13**(7): p. 293-301.
29. von Helmholtz, H., *Handuch der phsyiologik Optik*. 1867, Leipzig: Voss.
30. Gregory, R.L., *Perceptions as hypotheses.* Philos Trans R Soc Lond B Biol Sci, 1980. **290**(1038): p. 181-97.
31. Feldman, H. and K.J. Friston, *Attention, uncertainty, and free-energy.* Front Hum Neurosci, 2010. **4**: p. 215.
32. Friston, K.J., et al., *Action and behavior: a free-energy formulation.* Biological Cybernetics, 2010. **102**(3): p. 227-60.
33. Adams, R.A., S. Shipp, and K.J. Friston, *Predictions not commands: active inference in the motor system.* Brain Struct Funct, 2013. **218**(3): p. 611-43.
34. Petzschner, F.H., et al., *Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis.* Biol Psychiatry, 2017. **82**(6): p. 421-430.
35. Friston, K.J., et al., *Perceptions as hypotheses: saccades as experiments.* Frontiers in Psychology, 2012. **3**: p. 151.
36. Brown, H., et al., *Active inference, sensory attenuation and illusions.* Cogn Process, 2013. **14**(4): p. 411-27.
37. Stephan, K.E., et al., *Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression.* Front Hum Neurosci, 2016. **10**: p. 550.
38. Sterling, P., *Homeostasis vs allostasis: implications for brain function and mental disorders.* JAMA Psychiatry, 2014. **71**(10): p. 1192-3.
39. Ramsay, D.S. and S.C. Woods, *Clarifying the roles of homeostasis and allostasis in physiological regulation.* Psychol Rev, 2014. **121**(2): p. 225-47.
40. Sterling, P., *Allostasis: a model of predictive regulation.* Physiol Behav, 2012. **106**(1): p. 5-15.
41. Conant, R. and W.R. Ashby, *Every good regulator of a system must be a model of that system.* International Journal of Systems Science, 1970. **1**(2): p. 89-97.
42. Powers, W.T., *Behavior: The control of perception*. 1973, Hawthorne, NY: Aldine de Gruyter.
43. McLeod, P., N. Reed, and Z. Dienes, *Psychophysics: how fielders arrive in time to catch the ball.* Nature, 2003. **426**(6964): p. 244-5.
44. Friston, K.J., *The free-energy principle: a unified brain theory?* Nat Rev Neurosci, 2010. **11**(2): p. 127-38.
45. Chanes, L. and L.F. Barrett, *Redefining the Role of Limbic Areas in Cortical Processing.* Trends Cogn Sci, 2016. **20**(2): p. 96-106.
46. Critchley, H.D., C.J. Mathias, and R.J. Dolan, *Neuroanatomical basis for first- and second-order representations of bodily states.* Nat Neurosci, 2001. **4**(2): p. 207-12.
47. Barrett, L.F., K.S. Quigley, and P. Hamilton, *An active inference theory of allostasis and interoception in depression.* Philos Trans R Soc Lond B Biol Sci, 2016. **371**(1708).
48. Parr, T. and K.J. Friston, *Uncertainty, epistemics and active inference.* J R Soc Interface, 2017. **14**(136).
49. Friston, K., et al., *Active inference and epistemic value.* Cogn Neurosci, 2015. **6**(4): p. 187-214.

50. Bongard, J., V. Zykov, and H. Lipson, *Resilient machines through continuous self-modeling.* Science, 2006. **314**(5802): p. 1118-21.
51. Dayan, P. and N.D. Daw, *Decision theory, reinforcement learning, and the brain.* Cogn Affect Behav Neurosci, 2008. **8**(4): p. 429-53.
52. Allen, M. and K.J. Friston, *From cognitivism to autopoiesis: towards a computational framework for the embodied mind.* Synthese, 2016.
53. Baltieri, M. and C. Buckley, *An active inference implementation of phototaxis.* arXiv, 2017: p. 1707.01806
54. Noe, A., *Action in perception*. 2004, Cambridge, MA: MIT Press.
55. Seth, A.K., *A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia.* Cogn Neurosci, 2014. **5**(2): p. 97-118.
56. Seth, A.K., *Presence, objecthood, and the phenomenology of predictive perception.* Cognitive Neuroscience, 2015. **7**: p. 1-7.
57. Blanke, O., M. Slater, and A. Serino, *Behavioral, Neural, and Computational Principles of Bodily Self-Consciousness.* Neuron, 2015. **88**(1): p. 145-66.
58. Tsakiris, M., *My body in the brain: a neurocognitive model of body-ownership.* Neuropsychologia, 2010. **48**(3): p. 703-12.
59. Aspell, J.E., et al., *Turning the body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception.* Psychological Science, 2013. **24**(12): p. 2445-53.
60. Suzuki, K., et al., *Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion.* Neuropsychologia, 2013. **51**(13): p. 2909-17.
61. James, W., *The principles of psychology*. 1890, New York: Henry Holt.
62. Limanowski, J. and K. Friston, *'Seeing in the dark': Grounding phenomenal transparency and opacity in precision estimation for active inference.* Front Psychol, 2018. **9**: p. 643.
63. O'Regan, J.K. and A. Noë, *A sensorimotor account of vision and visual consciousness.* Behav Brain Sci, 2001. **24**(5): p. 939-73; discussion 973-1031.
64. Tsakiris, M., *The multisensory basis of the self: From body to identity to others [Formula: see text].* Q J Exp Psychol (Hove), 2017. **70**(4): p. 597-609.
65. Gibson, J.J., *The ecological approach to visual perception*. 1979, Hillsdale, NJ: Lawrence Erlbaum.
66. Ashby, W.R., *Design for a brain*. 1952, London, UK: Chapman and Hall.
67. Buckley, C., et al., *The free energy principle for action and perception: A mathematical review.* Journal of Mathematical Psychology, 2017. **81**: p. 55-79.
68. Knill, D.C. and A. Pouget, *The Bayesian brain: the role of uncertainty in neural coding and computation.* Trends Neurosci, 2004. **27**(12): p. 712-9.
69. Marr, D., *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 1982, New York: Freeman.
70. Friston, K.J., *A theory of cortical responses.* Philos Trans R Soc Lond B Biol Sci, 2005. **360**(1456): p. 815-36.
71. Godfrey-Smith, P.G., *Spencer and Dewey on life and mind*, in *The philosophy of artificial life*, M. Boden, Editor. 1996, Oxford University Press: Oxford. p. 314-331.
72. Makari, G., *Soul machine: The invention of the modern mind*. 2016, London: W. W. Norton.
73. de La Mettrie, J.O., *L'homme machine*. 1748, Leiden: Luzac.
74. Rosenfield, L.C., *From beast-machine to man-machine: Animal soul in French letters from Descartes to La Mettrie. New and enlarged edition*. 1968, New York: Octagon Books.
75. Putnam, H., *Representation and reality*. 1988, Cambridge, MA: MIT Press.
76. Seth, A.K. *The real problem*. Aeon 2016.

77. Schachter, S. and J.E. Singer, *Cognitive, social, and physiological determinants of emotional state.* Psychol Rev, 1962. **69**: p. 379-99.
78. Tracy, J.L. and D. Randles, *Four models of basic emotions: A review of Ekman and Cordado, Izard, Levenson, and Panksepp and Watt.* Emotion Review, 2011. **3**(4): p. 397-405.
79. Barrett, L.F. and A.B. Satpute, Neuroscience Letters, 2017.
80. Pinto, Y., et al., *Expectations accelerate entry of visual stimuli into awareness.* J Vis, 2015. **15**(8): p. 13.
81. Powers, A.R., C. Mathys, and P.R. Corlett, *Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors.* Science, 2017. **357**(6351): p. 596-600.
82. Joffily, M. and G. Coricelli, *Emotional valence and the free-energy principle.* PLoS Comput Biol, 2013. **9**(6): p. e1003094.
83. Clark, J.E., S. Watson, and K.J. Friston, *What is mood? A computational perspective.* Psychol Med, 2018: p. 1-8.
84. Petzschner, F.H., et al., *Focus of attention modulates the heartbeat evoked potential.* bioRxiv, 2018.
85. Kirchhoff, M., et al., *The Markov blankets of life: autonomy, active inference and the free energy principle.* J R Soc Interface, 2018. **15**(138).
86. Friston, K., et al., *Active Inference: A Process Theory.* Neural Comput, 2017. **29**(1): p. 1-49.
87. Owens, A.P., et al., *Interoceptive inference: From computational neuroscience to clinic.* Neurosci Biobehav Rev, 2018. **90**: p. 174-183.
88. Paulus, M.P. and M.B. Stein, *An insular view of anxiety.* Biological psychiatry, 2006. **60**(4): p. 383-7.
89. Sierra, M. and A.S. David, *Depersonalization: a selective impairment of self-awareness.* Consciousness and cognition, 2011. **20**(1): p. 99-108.
90. Quattrocki, E. and K. Friston, *Autism, oxytocin and interoception.* Neurosci Biobehav Rev, 2014. **47C**: p. 410-430.
91. Palmer, C.J., A.K. Seth, and J. Hohwy, *The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism.* Conscious Cogn, 2015. **36**: p. 376-89.
92. Ondobaka, S., J. Kilner, and K. Friston, *The role of interoceptive inference in theory of mind.* Brain Cogn, 2017. **112**: p. 64-68.
93. Baron-Cohen, S., A.M. Leslie, and U. Frith, *Does the autistic child have a "theory of mind"?* Cognition, 1985. **21**(1): p. 37-46.
94. Paton, B., J. Hohwy, and P.G. Enticott, *The rubber hand illusion reveals proprioceptive and sensorimotor differences in autism spectrum disorders.* J Autism Dev Disord, 2012. **42**(9): p. 1870-83.
95. Garfinkel, S.N., et al., *Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety.* Biol Psychol, 2016. **114**: p. 117-26.
96. Abramson, L.Y., M.E. Seligman, and J.D. Teasdale, *Learned helplessness in humans: critique and reformulation.* J Abnorm Psychol, 1978. **87**(1): p. 49-74.
97. Corlett, P.R. and P.C. Fletcher, *Computational psychiatry: a Rosetta Stone linking the brain to mental illness.* Lancet Psychiatry, 2014. **1**(5): p. 399-402.
98. Young, A.W. and K.M. Leafhead, *Betwixt life and death: case studies of the Cotard delusion*, in *Method in Madness*, P.W. Halligan and J.C. Marshall, Editors. 1996, Psychology Press: Hove, UK.
99. Brooks, R., *Intelligence without representation.* Artificial Intelligence, 1991. **47**: p. 139-160.

100. Fox, N.A., et al., *Developmental psychophysiology: Conceptual and methodological issues*, in *Handbook of psychophysiology*, J.T. Cacciopo, L.G. Tassinary, and G.G. Berntson, Editors. 2007, Cambridge University Press: New York, NY. p. 453-481.
101. Rinaman, L. and T.J. Koehnle, *The development of central visceral circuits*, in *Oxford Handbook of Developmental Behavioral Neuroscience*, M.S. Blumberg, J.H. Freeman, and S.R. Robinson, Editors. 2009, Oxford University Press: Oxford. p. 298-322.
102. McLaughlin, K.A., et al., *Causal effects of the early caregiving environment on development of stress response systems in children.* Proc Natl Acad Sci U S A, 2015. **112**(18): p. 5637-42.
103. Feldman, R., et al., *Mother and infant coordinate heart rhythms through episodes of interaction synchrony.* Infant Behav Dev, 2011. **34**(4): p. 569-77.
104. Fracasso, M.P., et al., *Cardiac activity in infancy: Reliability and stability of individual differences.* Infant Behavior & Development, 1994. **17**(3): p. 277-284.
105. Atzil, S. and L.F. Barrett, *Social regulation of allostasis: Commentary on "Mentalizing homeostasis: The social origins of interoceptive inference" by Fotopoulo and Tsakiris.* Neuropsychoanalysis, 2017. **19**(1): p. 29-33.
106. Abraham, E., et al., *Interoception sensitivity in the parental brain during the first months of parenting modulates children's somatic symptoms six years later: The role of oxytocin.* Int J Psychophysiol, 2018.
107. Goupil, L. and S. Kouider, *Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants.* Curr Biol, 2016. **26**(22): p. 3038-3045.
108. Brewer, R., et al., *Commentary on "Autism, oxytocin and interoception": Alexithymia, not Autism Spectrum Disorders, is the consequence of interoceptive failure.* Neurosci Biobehav Rev, 2015. **56**: p. 348-53.
109. Badoud, D. and M. Tsakiris, *From the body's viscera to the body's image: Is there a link between interoception and body image concerns?* Neurosci Biobehav Rev, 2017. **77**: p. 237-246.