The Cybernetic Bayesian Brain

From Interoceptive Inference to Sensorimotor Contingencies

Anil K. Seth

Is there a single principle by which neural operations can account for perception, cognition, action, and even consciousness? A strong candidate is now taking shape in the form of "predictive processing". On this theory, brains engage in predictive inference on the causes of sensory inputs by continuous minimization of prediction errors or informational "free energy". Predictive processing can account, supposedly, not only for perception, but also for action and for the essential contribution of the body and environment in structuring sensorimotor interactions. In this paper I draw together some recent developments within predictive processing that involve predictive modelling of internal physiological states (*interoceptive inference*), and integration with "enactive" and "embodied" approaches to cognitive science (*predictive perception of sensorimotor contingencies*). The upshot is a development of predictive processing that originates, not in Helmholtzian perception-as-inference, but rather in 20th-century cybernetic principles that emphasized homeostasis and predictive control. This way of thinking leads to (i) a new view of emotion as active interoceptive inference; (ii) a common predictive framework linking experiences of body ownership, emotion, and exteroceptive perception; (iii) distinct interpretations of active inference as involving disruptive and disambiguatory—not just confirmatory—actions to test perceptual hypotheses; (iv) a neurocognitive operationalization of the "mastery of sensorimotor contingencies" (where sensorimotor contingencies reflect the rules governing sensory changes produced by various actions); and (v) an account of the sense of subjective reality of perceptual contents ("perceptual presence") in terms of the extent to which predictive models encode potential sensorimotor relations (this being "counterfactual richness"). This is rich and varied territory, and surveying its landmarks emphasizes the need for experimental tests of its key contributions.

Keywords
Active inference | Counterfactually-equipped predictive model | Evolutionary robotics | Free energy principle | Interoception | Perceptual presence | Predictive processing | Sensorimotor contingencies | Somatic marker hypothesis | Synaesthesia

Author

Anil K. Seth
a.k.seth@sussex.ac.uk
University of Sussex
Brighton, United Kingdom

Commentator

Wanja Wiese
wawiese@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

An increasingly popular theory in cognitive science claims that brains are essentially prediction machines (Hohwy 2013). The theory is variously known as the Bayesian brain (Knill & Pouget 2004; Pouget et al. 2013), predictive processing (Clark 2013; Clark this collection), and the predictive mind (Hohwy 2013; Hohwy this collection), among others; here we use the term PP (predictive processing). (See Table 1 for a glossary of technical terms.) At its most fundamental, PP says that perception is the result of the brain inferring the most likely causes of its sensory inputs by minimizing the difference between actual sensory signals and the signals expected on the basis of continuously updated predictive models. Arguably, PP provides the most complete framework to date for explaining perception, cognition, and action in terms of fundamental theoretical principles and neurocognitive architectures. In this paper I describe a version of PP that is distinguished by (i) an emphasis on predictive modelling of in-

Table 1: A glossary of technical terms.

Allostasis	The process of achieving homeostasis.
Active inference	Classically conceived as the minimization of prediction error by performing actions that confirm sensory predictions. However, as argued in this paper, it may also involve the performance of actions to disconfirm current predictions or to disambiguate among competing perceptual hypotheses.
Counterfactually-equipped predictive model	A predictive or generative model that encodes not only the likely causes of current sensory inputs but also (and explicitly) the likely causes of fictive sensory inputs conditioned on possible but unexecuted actions.
Counterfactual richness	A predictive model is counterfactually rich if it encodes a rich repertoire of potential sensorimotor relations, i.e., relations between potential actions and their expected sensory consequences.
Exteroception/exteroceptive	The classic senses conveying signals originating in the external environment.
Free energy	An information-theoretic quantity that bounds or limits the surprise associated with encountering an input, given a generative/predictive model mapping causes to sensory inputs. Under fairly general assumptions, free energy is the long-run sum of prediction error.
Free energy principle (FEP)	The FEP says that organisms obey a fundamental imperative towards the avoidance of (information-theoretically) surprising events, according to which they must minimize the long-run average surprise of sensory states, since surprising sensory states are (in the long run) likely to reflect conditions incompatible with continued existence.
Homeostasis	Any regulative processes that enable a system to keep certain variables within specific bounds.
Interoception/interoceptive	The sense of the internal physiological condition of the body.
Interoceptive inference	The predictive modelling of internal physiological states.
Interoceptive sensitivity	A characterological trait that reflects individual sensitivity to interoceptive signals, usually operationalized via heartbeat detection tasks.
Perceptual presence	The sense of the subjective reality of the contents of perception.
PPSMC	Predictive Perception of SensoriMotor Contingencies. A new theory that integrates predictive processing with sensorimotor theory. It says that mastery of a sensorimotor contingency is equivalent to the induction and deployment of a counterfactually-equipped predictive model linking potential actions to their expected sensory consequences.
Predictive processing (PP)/predictive coding	A scheme, dating back at least to Hermann von Helmholtz, which conceives of perception as probabilistic inference on the causes of sensory signals. Predictive coding is one specific implementation of predictive processing that rests on algorithms developed in the setting of data compression.
Sensorimotor contingency (SMC)	SMCs describe ways in which sensory signals change given actions in specific contexts; they are "rules" describing sensorimotor dependencies.
Sensorimotor theory	A cognitive theory which says that visual experiences arises from an implicit knowledge or mastery of SMCs. On this theory, perception is an activity.

ternal physiological states and (ii) engagement with alternative frameworks under the banner of "enactive" and "embodied" cognitive science (Varela et al. 1993).

I first identify an unusual starting point for PP, not in Helmholtzian perception-as-inference, but in the mid 20th-century cybernetic theories associated with W. Ross Ashby (1952, 1956; Conant & Ashby 1970). Linking these origins to their modern expression in Karl Friston's "free energy principle" (2010), perception emerges as a *consequence* of a more fundamental imperative towards homeostasis and control, and not as a process designed to furnish a detailed inner "world model" suitable for cognition and action planning. The ensuing view of PP, while still fluently accounting for (exteroceptive) perception, turns out to be more naturally applicable to the predictive perception of internal bodily states, instantiating a process of *interoceptive inference* (Seth 2013; Seth et al. 2011). This concept provides a natural way of thinking of the neural substrates of emotional and mood experiences, and also describes a common mechanism by which interoceptive and exteroceptive signals can be integrated to provide a unified experience of body ownership and conscious selfhood (Blanke & Metzinger 2009; Limanowski & Blankenburg 2013).

The focus on embodiment leads to distinct interpretations of *active inference*, which in general refers to the selective sampling of sensory signals so as to improve perceptual predictions. The simplest interpretation of active inference is the changing of sensory data (via selective sampling) to conform to current predictions (Friston et al. 2010). However, by analogy with hypothesis testing in science, active inference can also involve seeking evidence that goes *against* current predictions, or that *disambiguates* multiple competing hypotheses. A nice example of the latter comes from self-modelling in evolutionary robotics, where multiple competing self-models are used to specify actions that are most likely to provide disambiguatory sensory evidence (Bongard et al. 2006). I will spend more time on this example later. Crucially, these different senses of active inference rest on the capacity of predictive models to encode *counterfactual* relations linking potential (but not necessarily executed) actions to their expected sensory consequences (Friston et al. 2012; Seth 2014b). It also implies the involvement of model comparison and selection—not just the optimization of parameters assuming a single model. These points represent significant developments in the basic infrastructure of PP.

The notion of counterfactual predictions connects PP with what at first glance seems to be its natural opponent: "enactive" theories of perception and cognition that explicitly reject internal models or representations (Clark this collection; Hutto & Myin 2013; Thompson & Varela 2001). Central to the enactive approach are notions of "sensorimotor contingencies" and their "mastery" (O'Regan & Noë 2001), where a sensorimotor contingency refers to a rule governing how sensory signals change in response to action. On this approach, the perceptual experience of (for example) redness is given by an implicit knowledge (mastery) of the way red things behave given certain patterns of sensorimotor activity. This mastery of sensorimotor contingencies is also said to underpin *perceptual presence*: the sense of subjective reality of the contents of perception (Noë 2006). From the perspective of PP, mastery of a sensorimotor contingency corresponds to the learning of a counterfactually-equipped predictive model connecting potential actions to expected sensory consequences. The resulting theory of PPSMC (Predictive Perception of SensoriMotor Contingencies), Seth 2014b) provides a much needed reconciliation of enactive and predictive theories of perception and action. It also provides a solution to the challenge of perceptual presence within the setting of PP: perceptual presence obtains when the underlying predictive models are *counterfactually rich*, in the sense of encoding a rich repertoire of potential (but not necessarily executed) sensorimotor relations. This approach also helps explain instances where perceptual presence seems to be lacking, such as in synaesthesia.

This is both a conceptual and theoretical paper. Space limitations preclude any significant treatment of the relevant experimental lit-

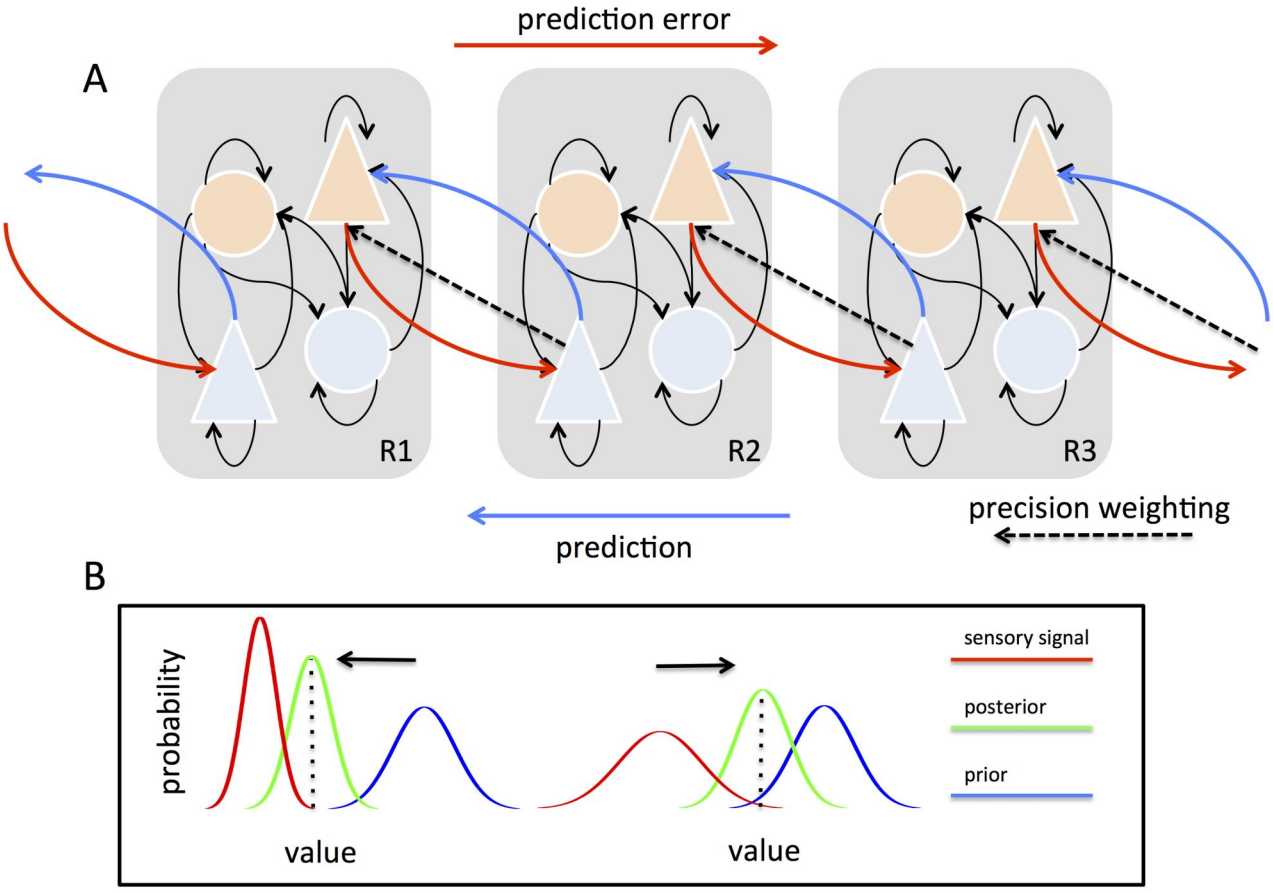

Figure 1: A. Schemas of hierarchical predictive coding across three cortical regions; the lowest on the left (R1) and the highest on the right (R3). Bottom-up projections (red) originate from "error units" (orange) in superficial cortical layers and terminate on "state units" (light blue) in the deep (infragranular) layers of their targets; while top-down projections (dark blue) convey predictions originating in deep layers and project to the superficial layers of their targets. Prediction errors are associated with precisions, which determine the relative influence of bottom-up and top-down signal flow via precision weighting (dashed lines). **B.** The influence of precisions on Bayesian inference and predictive coding. The curves show 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 sensory precision weighting has the opposite effect. Figure adapted from Seth (2013).

erature. However, even an exhaustive treatment would reveal that this literature so far provides only circumstantial support for the basics of PP, let alone for the extensions described here. Yet an advantage of PP theories is that they are grounded in concrete computational processes and neurocognitive architectures, giving us confidence that informative experimental tests can be devised. Implementing such an experimental agenda stands as a critical challenge for the future.

2 The predictive brain and its cybernetic origins

2.1 Predictive processing: The basics

PP starts with the assumption that in order to support adaptive responses, the brain must discover information about the external "hidden" causes of sensory signals. It lacks any direct access to these causes, and can only use information found in the flux of sensory signals them-

selves. According to PP, brains meet this challenge by attempting to predict sensory inputs on the basis of their own emerging models of the causes of these inputs, with prediction errors being used to update these models so as to minimize discrepancies. The idea is that a brain operating this way will come to encode (in the form of predictive or generative models) a rich body of information about the sources of signals by which it is regularly perturbed (Clark 2013).

Applied to cortical hierarchies, PP overturns classical notions of perception that describe a largely "bottom-up" process of evidence accumulation or feature detection. Instead, PP proposes that perceptual content is determined by top-down predictive signals emerging from multi-layered and hierarchically-organized generative models of the causes of sensory signals (Lee & Mumford 2003). These models are continually refined by mismatches (prediction errors) between predicted signals and actual signals across hierarchical levels, which iteratively update predictive models via approximations to Bayesian inference (see Figure 1). This means that the brain can induce accurate generative models of environmental hidden causes by operating only on signals to which it has direct access: *predictions* and *prediction errors*. It also means that even low-level perceptual content is determined via cascades of predictions flowing from very general abstract expectations, which constrain successively more fine-grained predictions.

Two further aspects of PP need to be emphasized from the outset. First, sensory prediction errors can be minimized either "passively", by changing predictive models to fit incoming data (perceptual inference), or "actively", by performing actions to confirm or test sensory predictions (active inference). In most cases these processes are assumed to unfold continuously and simultaneously, underlining a deep continuity between perception and action (Friston et al. 2010; Verschure et al. 2003). This process of active inference will play a key role in much of what follows. Second, predictions and prediction errors in a Bayesian framework have associated *precisions* (inverse variances, Figure 1). The precision of a prediction error is an indicator of its reliability, and hence can be used to determine its influence in updating top-down predictive models. Precisions, like mean values, are not given but must be inferred on the basis of top-down models and incoming data; so PP requires that agents have *expectations about precisions* that are themselves updated as new data arrive (and new precisions can be estimated). Precision expectations can therefore balance the influence of different prediction-error sources on the updating of predictive models. And if prediction errors have low (expected) precision, predictive models may overwhelm error signals (hallucination) or elicit actions that confirm sensory predictions (active inference).

A picture emerges in which cortical networks engage in recurrent interactions whereby bottom-up prediction errors are continuously reconciled with top-down predictions at multiple hierarchical levels—a process modulated at all times by precision weighting. The result is a brain that not only encodes information about the sources of signals that impinge upon its sensory surfaces, but that also encodes information about how its own actions interact with these sources in specifying sensory signals. *Perception* involves updating the parameters of the model to fit the data; *action* involves changing sensory data to fit (or test) the model; and *attention* corresponds to optimizing model updating by giving preference to sensory data that are expected to carry more information, which is called precision weighting (Hohwy 2013). This view of the brain is shamelessly model-based and representational (though with a finessed notion of representation), yet it also deeply embeds the close coupling of perception and action and, as we will see, the importance of the body in the mediation of this interaction.

2.2 Predictive processing and the free energy principle

PP can be considered a special case of the *free energy principle*, according to which perceptual inference and action emerge as a consequence of a more fundamental imperative towards the avoidance of "surprising" events (Friston 2005, 2009, 2010). On the free energy principle, or-

ganisms – by dint of their continued survival—must minimize the long-run average surprise of sensory states, since surprising sensory states are likely to reflect conditions incompatible with continued existence (think of a fish out of water). "Surprise" is not used here in the psychological sense, but in an information-theoretic sense—as the negative log probability of an event's occurrence (roughly, the unlikeliness of the occurrence of an event).

The connection with PP arises because agents cannot directly evaluate the (information-theoretic) surprise associated with an event, since this would require—impossibly—the agent to average over all possible occurrences of the event in all possible situations. Instead, the agent can only maintain a lower limit on surprise by minimizing the difference between actual sensory signals and those signals predicted according to a generative or predictive model. This difference is *free energy*, which, under fairly general assumptions, is the long-run sum of prediction error.

An attractive feature of the free energy principle is that it brings to the table a rich mathematical framework that shows how PP can work in practice. Formally, PP depends on established principles of Bayesian inference and model specification, whereby the most likely causes of observed data (*posterior*) are estimated based on optimally combining *prior expectations* of these causes with observed data, by using a (generative, predictive) model of the data that would be observed given a particular set of causes (*likelihood*). (See Figure 1 for an example of priors and posteriors.) In practice, because optimal Bayesian inference is usually intractable, a variety of approximate methods can be applied (Hinton & Dayan 1996; Neal & Hinton 1998). Friston's framework appeals to previously worked-out "variational" methods, which take advantage of certain approximations (e.g., Gaussianity, independence of temporal scales)—thus allowing a potentially neat mapping onto neurobiological quantities (Friston et al. 2006).[1]

1 Some challenging questions surface here as to whether prediction errors are used to update priors, which corresponds to standard Bayesian inference, or whether they are used to update the underlying generative/predictive model, which corresponds to learning.

The free energy principle also emphasizes *action* as a means of prediction error minimization, this being *active inference*. In general, active inference involves the selective sampling of sensory signals so as to minimize uncertainty in perceptual hypotheses (minimizing the entropy of the posterior). In one sense this means that actions are selected to provide evidence compatible with current perceptual predictions. This is the most standard interpretation of the concept, since it corresponds most directly to minimization of prediction error (Friston 2009). However, as we will see, actions can also be selected on the basis of an attempt to find evidence going against current hypotheses, and/or to efficiently disambiguate between competing hypotheses. These finessed senses of active inference represent developments of the free energy framework. Importantly, action itself can be thought of as being brought about by the minimization of *proprioceptive* prediction errors via the engagement of classical reflex arcs (Adams et al. 2013; Friston et al. 2010). This requires transiently low precision-weighting of these errors (or else predictions would simply be updated instead), which is compatible with evidence showing sensory attenuation during self-generated movements (Brown et al. 2013).

A more controversial aspect of the free energy principle is its claimed generality (Hohwy this collection). At least as described by Friston, it claims to account for adaptation at almost any granularity of time and space, from macroscopic trends in evolution, through development and maturation, to signalling in neuronal hierarchies (Friston 2010). However, in some of these interpretations reliance on predictive modelling is only implicit; for example the body of a fish can be considered to be an implicit model of the fluid dynamics and other affordances of its watery environment (see section 2.3). I am not concerned here with these broader interpretations, but will focus on those cases in which biological (neural) mechanisms plausibly implement explicit predictive inference via approximations to Bayesian computations—namely, the Bayesian brain (Knill & Pouget 2004; Pouget et al. 2013). Here, the free energy principle has potentially the greatest explanat-

ory power, especially given the convergence of empirical evidence (see Clark 2013 and Hohwy 2013 for reviews) and computational modelling showing how cortical microcircuits might implement approximate Bayesian inference (Bastos et al. 2012).

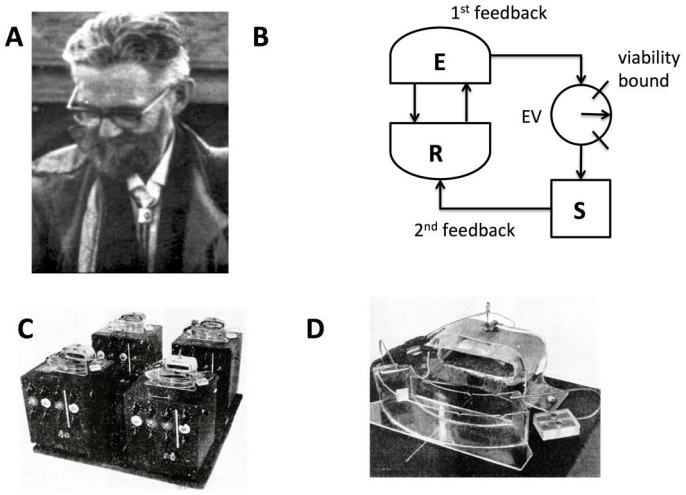

Figure 2: A. W. Ross Ashby, British psychiatrist and pioneer of cybernetics (1903–1972). **B.** A schematic of ultrastability, based on Ashby's notebooks. The system *R* homeostatically maintains its essential variables (EVs) within viability limits via first-order feedback with the environment *E*. When first-order feedback fails, so that EVs run out-of-bounds, second order "ultrastable" feedback is triggered so that *S* (an internal controller, potentially model-based) changes the parameters of R governing the first-order feedback. *S* continually changes *R* until homeostatic relations are regained, leaving the EVs again within bounds. **C.** Ashby's "homeostat", consisting of four interconnected ultrastable systems, forming a so-called "multistable" system. **D.** One ultrastable unit from the homeostat. Each unit had a trough of water with an electric field gradient and a metal needle. Instability was represented by the non-central needle positions, which on occurring would alter the resistances connecting the units via discharge through capacitors. For more details see Ashby (1952) and Pickering (2010).

2.3 Predictive processing, free energy, and cybernetics

Typically, the origins of PP are traced to the work of the 19th Century physiologist Hermann von Helmholtz, who first formalized the idea of perception as inference. However, the Helmholtzian view is rather passive, inasmuch as there is little discussion of active inference or behaviour. The close coupling of perception and action emphasized in the free energy principle points instead to a deep connection between PP and mid-twentieth-century cybernetics. This is most obvious in the works of W. Ross Ashby (Ashby 1952; 1956; Conant & Ashby 1970) but is also evident more generally (Dupuy 2009; Pickering 2010). Importantly, cybernetics adopted as its central focus the *prediction and control of behaviour* in so-called teleological or purposeful machines.[2] More precisely, cybernetic theorists were (are) interested in systems that appear to have goals (i.e., teleological) and that participate in circular causal chains (i.e., involving feedback) coupling goal-directed sensation and action.

Two key insights from the first wave of cybernetics usefully anticipate the core developments of PP within cognitive science. These are both associated with Ashby, a key figure in the movement and often considered its leader, at least outside the USA (Figure 2).

The first insight consists in an emphasis on the homeostasis of internal *essential variables*, which, in physiological settings, correspond to quantities like blood pressure, heart rate, blood sugar levels, and the like. In Ashby's framework, when essential variables move beyond specific viability limits, adaptive processes are triggered that re-parameterize the system until it reaches a new equilibrium in which homeostasis is restored (Ashby 1952). Such systems are, in Ashby's terminology, *ultrastable*, since they embody (at least) two levels of feedback: a first-order feedback that homeostatically regulates essential variables (like a thermostat) and a second-order feedback that allostatically[3] re-organises a system's input–output relations when first-order feedback fails, until a new homeostatic regime is attained. In the most basic case, as implemented in Ashby's famous "homeostat" (Figure 2), this second-order feedback simply involves random changes to system

2 This underlines the close links between cybernetics and behaviourism. Perhaps this explains why cybernetics was so reluctant to bring phenomenology into its remit, an exclusion which, looking back, seems like a missed opportunity.

3 Allostasis: the process of achieving homeostasis.

parameters until a new stable regime is reached. The importance of this insight for PP is that it locates the function of biological and cognitive processes in generalizing homeostasis to ensure that internal essential variables remain within expected ranges.

Another way to summarize the fundamental cybernetic principle is to say that adaptive systems ensure their continued existence by successfully responding to environmental perturbations so as to maintain their internal organization. This leads to the second insight, evident in Ashby's *law of requisite variety*. This states that a successful control system must be capable of entering at least as many states as the system being controlled: "only variety can force down variety" (Ashby 1956). This induces a functional boundary between controller and environment and implies a minimum level of complexity for a successful controller, which is determined by the causal complexity of the environmental states that constitute potential perturbations to a system's essential variables. This view was refined some years later, in a 1970 paper written with Roger Conant entitled "Every good regulator of a system must be a model of that system" (Conant & Ashby 1970). This paper builds on the law of requisite variety by arguing (and attempting to formally show) that the nature of a controller capable of suppressing perturbations imposed by an external system (e.g., the world) must instantiate a model of that system. This provides a clear connection with the free energy principle, which proposes that adaptive systems minimize a limit on free energy (long-run average surprise) by inducing and refining a generative model of the causes of sensory signals. It also moves beyond Ashby's homeostat by implying that model-based controllers can engage in more successful multilevel feedback than is possible by random variation of higher-order parameters.

Putting these insights together provides a distinctive way of seeing the relevance of PP to cognition and biological adaptation. It can be summarized as follows. The purpose of cognition (including perception and action) is to maintain the homeostasis of essential variables and of internal organization (ultrastability). This implies the existence of a control mechanism with sufficient complexity to respond to (i.e., suppress) the variety of perturbations it encounters (law of requisite variety). Further, this structure must instantiate a model of the system to be controlled (good regulator theorem), where the system includes both the body and the environment (and their interactions). As Ashby himself tells us "[t]he whole function of the brain can be summed up in: error correction" (quoted in Clark 2013, p. 1). Put this way, perception emerges as a *consequence* of a more fundamental imperative towards organizational homeostasis, and not as a stage in some process of internal world-model construction. This view, while highlighting different origins, closely parallels the assumptions of the free energy principle in proposing a primary imperative towards the continued survival of the organism (Friston 2010).

It may be surprising to consider the legacy of cybernetics in this light. This is because many previous discussions of this legacy focus on examples which show that complex, apparently goal-directed behaviour can emerge from simple mechanisms interacting with structured bodies and environments (Beer 2003; Braitenberg 1984). On this more standard development, cybernetics challenges rather than asserts the need for internal models and representations: it is often taken to justify slogans of the sort "the world is its own best model" (Brooks 1991). In fact, cybernetics is agnostic with respect to the need for deployment of explicit internally-specified predictive models. If environmental circumstances are reasonably stable, and mappings between perturbations and (homeostatic) responses reasonably straightforward, then the good regulator theorem can be satisfied by controllers that only implicitly model their environments. This is the case, for instance, in the Watt governor: a device that is able exquisitely to control the output of (for instance) a steam engine, in virtue of its mechanism, and not through the deployment of explicit predictive models or representations (see Figure 3 and Van Gelder 1995; note that the governor can

be described as an implicit model since it has variables – e.g., eccentricity of the metal balls from the central column – which map onto environmental variables that affect the homeostatic target – engine output). However, where there exist many-to-many mappings between sensory states and their probable causes, as may be the case more often than not, it will pay to engage explicit inferential processes in order to extract the most probable causes of sensory states, insofar as these causes threaten the homeostasis of essential variables.

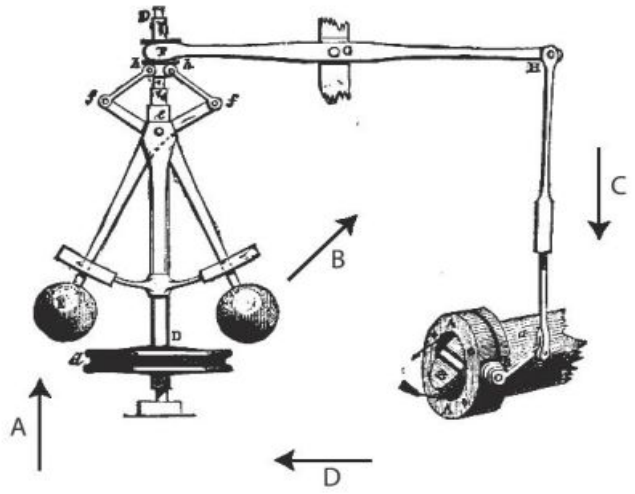

Figure 3: The Watt governor. This system, a central contributor to the industrial revolution, enabled precise control over the output of (for example) steam engines. As the speed of the engine increases, power is supplied to the governor (A) by a belt or chain, causing it to rotate more rapidly so that the metal balls have more kinetic energy. This causes the balls to rise (B), which closes the throttle valve (C), thereby reducing the steam flow, which in turn reduces engine speed (D). The opposite happens when the engine speed decreases, so that the governor maintains engine speed at a precise equilibrium.

In summary, rather than seeing PP as originating solely in the Helmholtzian notion of "perception as inference", it is fruitful to see it also as part of a process of model-based *predictive control* entailed by a fundamental imperative towards internal homeostasis. This shift in perspective reveals a distinctive agenda for PP in cognitive science, to which I shall now turn.

3 Interoceptive inference, emotion, and predictive selfhood

3.1 Interoceptive inference and emotion

Considering the cybernetic roots of PP, together with the free energy principle, leads to a potentially counterintuitive idea. This is that PP may apply more naturally to *interoception* (the sense of the internal physiological condition of the body) than to *exteroception* (the classic senses, which carry signals that originate in the external environment). This is because for an organism it is more important to avoid encountering unexpected interoceptive states than to avoid encountering unexpected exteroceptive states. A level of blood oxygenation or blood sugar that is unexpected is likely to be bad news for an organism, whereas unexpected exteroceptive sensations (like novel visual inputs) are less likely to be harmful and may in some cases be desirable, as organisms navigate a delicate balance between exploration and exploitation (Seth 2014a), testing current perceptual hypotheses through active inference (see section 5, below), all ultimately in the service of maintaining organismic homeostasis.

Perhaps because of its roots in Helmholtz, PP has largely been developed in the setting of visual neuroscience (Rao & Ballard 1999), with a related but somewhat independent line in motor control (Wolpert & Ghahramani 2000). Recently, an explicit application of PP to interoception has been developed (Seth 2013; Seth & Critchley 2013; Seth et al. 2011; see also Gu et al. 2013). On this theory of *interoceptive inference* (or equivalently *interoceptive predictive coding*), emotional states (i.e., subjective feeling states) arise from top-down predictive inference of the causes of interoceptive sensory signals (see Figure 4). In direct analogy to exteroceptive PP, emotional content is constitutively specified by the content of top-down interoceptive predictions *at a given time*, marking a distinction with the well-studied impact of expectations on *subsequent* emotional states (see e.g., Ploghaus et al. 1999; Ueda et al. 2003). Furthermore, interoceptive prediction errors can

be minimized by (i) updating predictive models (perception, corresponding to new emotional contents); (ii) changing interoceptive signals through engaging autonomic reflexes (autonomic control or active inference); or (iii) performing behaviour so as to alter external conditions that impact on internal homeostasis (allostasis; Gu & Fitzgerald 2014; Seth et al. 2011).

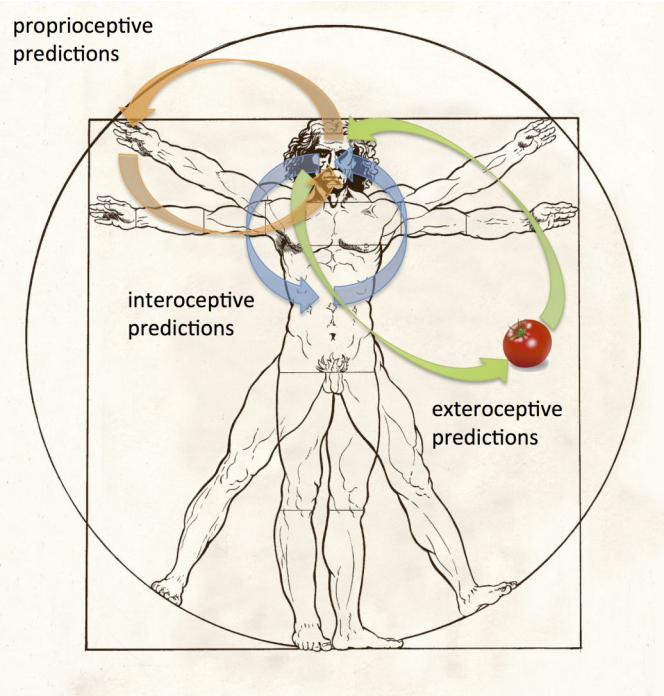

Figure 4: Inference and perception. Green arrows represent exteroceptive predictions and predictions errors underpinning perceptual content, such as the visual experience of a tomato. Orange arrows represent proprioceptive predictions (and prediction errors) underlying action and the experience of body ownership. Blue arrows represent interoceptive predictions (and prediction errors) underlying emotion, mood, and autonomic regulation. Hierarchically higher levels will deploy multimodal and even amodal predictive models spanning these domains, which are capable of generating multimodal predictions of afferent signals.

Consider an example in which blood sugar levels (an essential variable) fall towards or beyond viability thresholds, reaching unexpected and undesirable values (Gu & Fitzgerald 2014; Seth et al. 2011). Under interoceptive inference, the following responses ensue. First, interoceptive prediction error signals update top-down expectations, leading to subjective experiences of hunger or thirst (for sugary things). Because these feeling states are themselves surprising (and non-viable) in the long run, they signal prediction errors at hierarchically-higher levels, where predictive models integrate multimodal interoceptive and exteroceptive signals. These models instantiate predictions of temporal sequences of matched exteroceptive and interoceptive inputs, which flow down through the hierarchy. The resulting cascade of prediction errors can then be resolved either through autonomic control, in order to metabolize bodily fat stores (active inference), or through allostatic actions involving the external environment (i.e., finding and eating sugary things).

The sequencing and balance of these events is governed by relative precisions and their expectations. Initially, interoceptive prediction errors have high precision (weighting) given a higher-level expectation of stable homeostasis. Whether the resulting high-level prediction error engages autonomic control or allostatic behaviour (or both) depends on the precision weighting of the corresponding prediction errors. If food is readily available, consummatory actions lead to food intake (as described earlier, these actions are generated by the resolution of proprioceptive prediction errors). If not, autonomic reflexes initiate the metabolization of bodily fat stores, perhaps alongside appetitive behaviours that are predicted to lead to the availability of food, conditioned on performing these behaviours.[4]

3.2 Implications of interoceptive inference

Several interesting implications arise when considering emotion as resulting from interoceptive inference (Seth 2013). First, the theory generalizes previous "two factor" theories of emotion that see emotional content as resulting from an interaction between the perception of physiolo-

4 It is interesting to consider possible dysfunctions in this process. For example, if high-level predictions about the persistence of low blood sugar become abnormally strong (i.e., low blood sugar becomes chronically expected), allostatic food-seeking behaviours may not occur. This process, akin to the transition from hallucination to delusion in perceptual inference (Fletcher & Frith 2009), may help understand eating disorders in terms of dysfunctional signalling of satiety.

gical changes (James 1894) and "higher-level" cognitive appraisal of the context within which these changes occur (Schachter & Singer 1962). Instead of distinguishing "physiological" and "cognitive" levels of description, interoceptive inference sees emotional content as resulting from the multi-layered prediction of interoceptive input spanning many levels of abstraction. Thus, interoceptive inference integrates cognition and emotion within the powerful setting of PP.

The theory also connects with influential frameworks that link interoception with decision making, notably the "somatic marker hypothesis" proposed by Antonio Damasio (1994). According to the somatic marker hypothesis, intuitive decisions are shaped by interoceptive responses (somatic markers) to potential outcomes. This idea, when placed in the context of interoceptive inference, corresponds to the guidance of behavioural (allostatic) responses towards the resolution of interoceptive prediction error (Gu & Fitzgerald 2014; Seth 2014a). It follows that intuitive decisions should be affected by the degree to which an individual maintains accurate predictive models of his or her own interoceptive states; see Dunn et al. 2010, Sokol-Hessner et al. 2014 for evidence along these lines.

There are also important implications for disorders of emotion, selfhood, and decision-making. For example, anxiety may result from the chronic persistence of interoceptive prediction errors that resist top-down suppression (Paulus & Stein 2006). Dissociative disorders like alexithymia (the inability to describe one's own emotions), and depersonalization and derealisation (the loss of sense of reality of the self and world) may also result from dysfunctional interoceptive inference, perhaps manifest in abnormally low interoceptive precision expectations (Seth 2013; Seth et al. 2011). In terms of decision-making, it may be productive to think of addiction as resulting from dysfunctional active inference, whereby strong interoceptive priors are confirmed through action, overriding higher-order or hyper-priors relating to homeostasis and organismic integrity. It has even been suggested that autism spectrum disorders may originate in aberrant encoding of the salience or precision of interoceptive prediction errors (Quattrocki & Friston 2014). The reasoning here is that aberrant salience during development could disrupt the assimilation of interoceptive and exteroceptive cues within generative models of the "self", which would impair a child's ability to properly assign salience to socially relevant signals.

3.3 The predictive embodied self

The maintenance of physiological homeostasis solely through direct autonomic regulation is obviously limited: behavioural (allostatic) interactions with the world are necessary if the organism is to avoid surprising physiological states in the long run. The ability to deploy adaptive behavioural responses mandates the original Helmholtzian view of perception-as-inference, which has been the primary setting for the development of PP so far. A critical but arguably overlooked middle ground, which mediates between physiological state variables and the external environment, is the *body*. On one hand, the body is the material vehicle through which behaviour is expressed, permitting allostatic interactions to take place. On the other, the body is itself an essential part of the organismic system, the homeostatic integrity of which must be maintained. In addition, the experience of owning and identifying with a particular body is a key component of being a conscious self (Apps & Tsakiris 2014; Blanke & Metzinger 2009; Craig 2009; Limanowski & Blankenburg 2013; Seth 2013).

It is tempting to ask whether common predictive mechanisms could underlie not only classical exteroceptive perception (like vision) and interoception (see above), but also their integration in supporting conscious and unconscious representations of the body and self (Seth 2013). The significance of this question is underlined by realising that just as the brain has no direct access to causal structures in the external environment, it also lacks direct access to its own body. That is, given that the brain is in the business of inferring the causal sources of

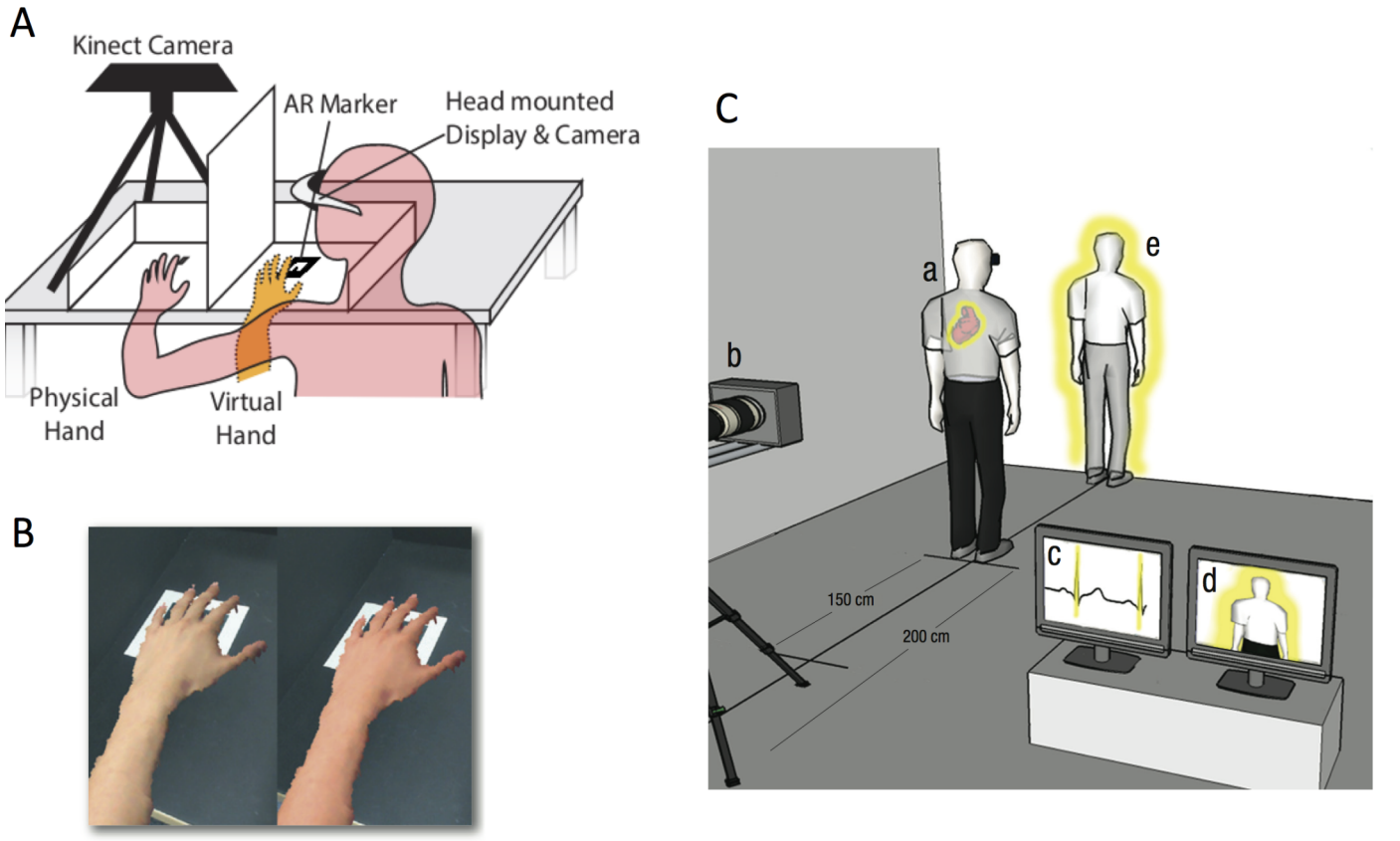

Figure 5: The interaction of interoceptive and exteroceptive signals in shaping the experience of body ownership. **A.** Set-up for applying cardio-visual feedback in the rubber hand illusion. A Microsoft Kinect obtains a real-time 3D model of a subject's left hand. This is re-projected into the subject's visual field using a head-mounted display and augmented reality (AR) software. **B.** The colour of the virtual hand is modulated by the subject's heart-beat. **C.** A similar set-up for the full-body illusion whereby a visual image of a subject's body is surrounded by a halo pulsing either in time or out of time with the heartbeat. Panels A and B are adapted from Suzuki et al. (2013); panel C is adapted from Aspell et al. (2013).

sensory signals, a key challenge emerges when distinguishing those signals that pertain to the body from those that originate from the external environment. A clue to how this challenge is met is that the physical body, unlike the external environment, constantly generates and receives internal input via its interoceptive and proprioceptive systems (Limanowski & Blankenburg 2013; Metzinger 2003). This suggests that the experienced body (and self) depends on the brain's best guess of the causes of those sensory signals most likely to be "me" (Apps & Tsakiris 2014), across interoceptive, proprioceptive, and exteroceptive domains (Figure 4).

There is now considerable evidence that the *experience of body ownership* is highly plastic and depends on the multisensory integration of body-related signals (Apps & Tsakiris 2014; Blanke & Metzinger 2009). One classic example is the *rubber hand illusion*, where the stroking of an artificial hand synchronously with a participant's real hand, while visual attention is focused on the artificial hand, leads to the experience that the artificial hand is somehow part of the body (Botvinick & Cohen 1998). According to current multisensory integration models, this change in the experience of body ownership is due to correlation between vision and touch overriding conflicting proprioceptive inputs (Makin et al. 2008). Through the lens of PP, this implies that prediction errors induced by multisensory conflicts will over time update self-related priors (Apps & Tsakiris 2014), with different signal sources (vision, touch, proprioception) each precision-weighted according to their expected reliability, and all in

the setting of strong prior expectations for correlated input.[5]

While the potential for exteroceptive multisensory integration to modulate the experience of body ownership has been extensively explored both for the ownership of body parts and for the experience of ownership of the body as a whole (for reviews, see Apps & Tsakiris 2014; Blanke & Metzinger 2009), only recently has attention been paid to interactions between interoceptive and exteroceptive signals. Initial evidence in this line of investigation was indirect, for example showing correlation between susceptibility to the rubber hand illusion and individual differences in the ability to perceive interoceptive signals ("interoceptive sensitivity", typically indexed by heartbeat detection tasks; Tsakiris et al. 2011). Other relevant studies have shown that body ownership illusions lead to temperature reductions in the corresponding body parts, perhaps reflecting altered active autonomic inference (Moseley et al. 2008; Salomon et al. 2013).

Emerging evidence now points more directly towards the predictive multisensory integration of interoceptive and exteroceptive signals in shaping the experience of body ownership. Two recent studies have taken advantage of so-called "cardio-visual synchrony" where virtual-reality representations of body parts (Suzuki et al. 2013) or the whole body (Aspell et al. 2013) are modulated by simultaneously recorded heartbeat signals, with the modulation either in-time or out-of-time with the actual heartbeat (Figure 5). These data suggest that statistical correlations between interoceptive (e.g., cardiac) and exteroceptive (e.g., visual) signals can lead to the updating of predictive models of self-related signals through (hierarchical) minimization of prediction error, just as happens for purely exteroceptive multisensory conflicts in the classic rubber hand illusion.

While these studies underline the plausibility of common predictive mechanisms underlying emotion, selfhood, and perception, many open questions nevertheless remain. A key challenge is to detail the underlying neural operations. Though a detailed analysis is beyond the scope of the present paper, it is worth noting that attention is increasingly focused on the insular cortex (especially its anterior parts) as a potential source of interoceptive predictions, and also as a comparator registering interoceptive prediction errors. The anterior insula has long been considered a major cortical locus for the integration of interoceptive and exteroceptive signals (Craig 2003; Singer et al. 2009); it is strongly implicated in interoceptive sensitivity (Critchley et al. 2004); it is sensitive to interoceptive prediction errors—at least in some contexts (Paulus & Stein 2006); and it has a high density of so-called "von Economo" neurons,[6] which have been frequently though circumstantially associated with consciousness and selfhood (Critchley & Seth 2012; Evrard et al. 2012).

3.4 Active inference, self-modeling, and evolutionary robotics

What role might *active* inference play in predictive self-modelling? Autonomic changes during illusions of body ownership (see above) are consistent with active inference; however they do not speak directly to its function. In the classic rubber hand illusion, hand or finger movements can be considered active inferential tests of self-related hypotheses. If these movements are not reflected in the "rubber hand", the illusion is destroyed—presumably because predicted visual signals are not confirmed (Apps & Tsakiris 2014). However, if hand movements are mapped to a virtual "rubber hand"—through clever use of virtual and augmented reality—the illusion is in fact strengthened, presumably because the multisensory correlation of peri-hand visual and proprioceptive signals constitutes a more stringent test of the perceptual hypothesis of ownership of the virtual hand (Suzuki et al. 2013). This introduces the idea that active inference is not simply about confirming sensory predictions but also involves seeking "disruptive" actions that are most informative with respect to testing current predictions,

5 Interestingly the expectation of perceptual correlations seems to be sufficient for inducing the rubber hand illusion (Ferri et al. 2013).

6 These are long-range projection neurons found selectively in hominid primates and certain other species.

and/or at disambiguating competing predictions (Gregory 1980). A nice example of how this happens in practice comes from *evolutionary robotics*[7]—which is obviously a very different literature, though one that inherits directly from the cybernetic tradition.

In a seminal 2006 study, Josh Bongard and colleagues described a four-legged "starfish" robot that engaged in a process much like active inference in order to model its own morphology so as to be able to control its movement and attain simple behavioural goals (Bongard et al. 2006). While there are important differences between evolutionary robotics and (active) Bayesian inference, there are also broad similarities; importantly, both can be cast in terms of model selection and optimization.

The basic cycle of events is shown in Figure 6. The robot itself is shown in the centre (A). The goal is to develop a controller capable of generating forward movement. The challenge is that the robot's morphology is unknown to the robot itself. The system starts with a range of (generic prior) potential self-models (B), here specified by various configurations of three-dimensional physics engines. The robot performs a series of initially random actions and evaluates its candidate self-models on their ability to predict the resulting proprioceptive afferent signals. Even though all initial models will be wrong, some may be better than others. The key step comes next. The robot evaluates new candidate actions *on the extent to which the current best self-models make different predictions as to their (proprioceptive) consequences*. These disambiguating actions are then performed, leading to a new ranking of self-models based on their success at proprioceptive prediction. This ranking, via the evolutionary robotics methods of mutation and replication, gives rise to a new population of candidate self-models. The upshot is that the system swiftly develops accurate self-models that can be used to generate controllers enabling movement (D). An interesting feature of this process is that it is highly resilient to unexpected perturbations. For instance, if a leg is removed then proprioceptive prediction errors will immediately ensue. As a result, the system will engage in another round of self-model evolution (including the co-specification of competing self-models and disambiguating actions) until a new, accurate, self-model is regained. This revised self-model can then be used to develop a new gait, allowing movement, even given the disrupted body (E, F).[8]

This study emphasizes that the operational criterion for a successful self-model is not so much its fidelity to the physical robot, but rather its ability to predict sensory inputs under a repertoire of actions. This underlines that predictive models are recruited for the control of behaviour (as cybernetics assumes) and not to furnish general-purpose representations of the world or the body.

The study also provides a concrete example of how actions can be performed, not to achieve some externally specified goal, but to permit inference about the system's own physical instantiation. Bayesian or not, this implies active inference. Indeed, perhaps its most important contribution is that it highlights how active inference can prescribe *disruptive* or *disambiguating* actions that generate sensory prediction errors under competing hypotheses, and not just actions that seek to confirm sensory predictions. This recalls models of attention based on maximisation of Bayesian surprise (Itti & Baldi 2009), and is equivalent to hypothesis testing in science, where the best experiments are those concocted on the basis of being most likely to falsify a given hypothesis (disruptive) or distinguish between competing hypotheses (disambiguating). It also implies that agents encode predictions about the likely sensory consequences of a range of potential actions, allowing the selection of those actions likely to be the most disruptive or disambiguating. This concept of a *counterfactually-equipped predictive model* bring us nicely to our next topic: so-called *enactive* cognitive science and its relation to PP.

7 Evolutionary robotics involves the use of population-based search procedures (genetic algorithms) to automatically specify control architectures (and/or morphologies) of mobile robots. For an excellent introduction see (Bongard 2013).

8 Videos showing the evolution of both gait and self-model are available from http://creativemachines.cornell.edu/emergent_self_models

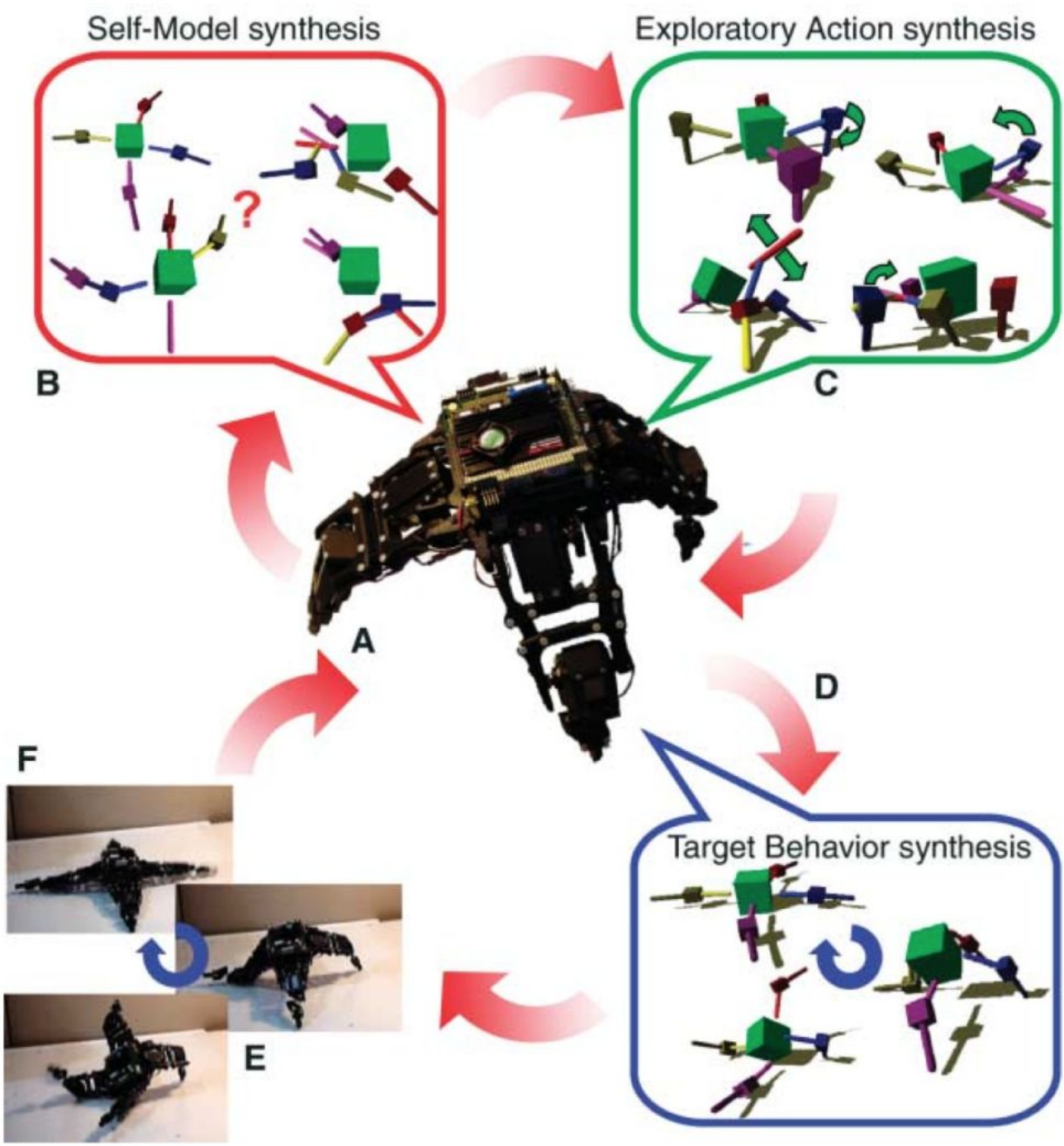

Figure 6: An evolutionary-robotics experiment demonstrating continuous self-modelling Bongard et al. (2006). See text for details. Reproduced with permission.

4 Predictive processing and enactive cognitive science

4.1 Enactive theories, weak and strong

The idea that the brain relies on internal representations or models of extra-cranial states of affairs has been treated with suspicion ever since the limitations of "good old fashioned artificial intelligence" became apparent (Brooks 1991). Many researchers of artificial intelligence have indeed returned to cybernetics as an alternative framework in which closely coupled feedback loops, leveraging invariants in brain-body-world interactions, obviate the need for detailed internal representations of external properties (Pfeifer & Scheier 1999). The evolutionary robotics methodology just described is

often coupled with simple dynamical neural networks in order to realize controllers that are tightly embodied and embedded in just this way (Beer 2003). Within cognitive science, such anti-representationalism is most vociferously defended by the movement variously known as "enactive" (Noë 2004), "embodied" (Gallese & Sinigaglia 2011), or "extended" (Clark & Chalmers 1998) cognitive science. Among these approaches, it is enactivism that is most explicitly anti-representationalist. While enactive theorists might agree that adaptive behaviour requires organisms and control structures that are systematically sensitive to statistical structures in their environment, most will deny that this sensitivity implies the existence and deployment of any "inner description" or model of these probabilistic patterns (Chemero 2009; Hutto & Myin 2013).

This tradition has weak and strong expressions. At the weak extreme is the truism that perception, cognition, and behaviour—and their underlying mechanisms—cannot be understood without a rich appreciation of the roles of the body, the environment, and the structured interactions that they support (Clark 1997; Varela et al. 1993). Weak enactivism is eminently compatible with PP, as seen especially with emerging versions of PP that stress embodiment through self-modelling and interoception, and which emphasize the importance of agent-environment coupling (embeddedness) through active inference. At the other extreme lie claims that explanations based on internal representations or models of any sort are fundamentally misguided, and that a new explicitly non-representational vocabulary is needed in order to make sense of the relations between brains, bodies, and the world (O'Regan et al. 2005). Strong enactivism is by definition incompatible with PP since it rejects the core concept of the internal model.

4.2 Sensorimotor contingency theory

A landmark in the strongly enactive approach is SMC (sensorimotor contingency) theory, which says that perception depends on the "practical mastery" of sensorimotor dependencies relevant to behaviour (O'Regan & Noë 2001). In brief, SMC theory claims that experience and perception are not things that are "generated" by the brain (or by anything else for that matter) but are, rather, "skills" consisting of fluid patterns of on-going interaction with the environment (O'Regan & Noë 2001). For instance, on SMC theory the conscious visual experience of redness is given by *the exercise of practical mastery of the laws governing how interactions with red things unfold* (these laws being the "SMC"s). The theory is not, however, limited to vision: the experiential quality of the softness of a sponge would be given by (practical mastery of) the laws governing its squishiness upon being pressed.

Two aspects of SMC theory deserve emphasis here. The first is that the concept of an SMC rightly underlines the close coupling of perception and action and the critical importance of ongoing agent-environment interaction in structuring perception, action, and behaviour. This is inherited from Gibsonian notions of perceptual affordance (Gibson 1979) and has certainly advanced our understanding of why different kinds of perceptual experience (vision, smell, touch, etc.) have different qualitative characters.

The second is that *mastery* of an SMC requires an essentially *counterfactual* knowledge of relations between particular actions and the resulting sensations. In vision, for instance, mastery entails an implicit knowledge of the ways in which moving our eyes and bodies would reveal additional sensory information about perceptual objects (O'Regan & Noë 2001). Here SMC theory has made an important contribution to our understanding of *perceptual presence*. Perceptual presence refers to the property whereby (in normal circumstances) perceptual contents appear as subjectively real, that is, as *existing*. For example, when viewing a tomato, we see it as real inasmuch as we seem to be perceptually aware of some of its parts (e.g., its back) that are not currently causally impacting our sensory surfaces. Looking at a picture of a tomato does not give rise to the same subjective impression of realness. But how can we be aware of parts of the tomato that, strictly speaking, we do not

see? SMC theory says the answer lies in our (implicit) mastery of SMCs, which relate potential actions to their likely sensory effects; and it is in this sense that we can be perceptually aware of parts of the tomato that we cannot actually see (Noë 2006).

SMC theory has often been set against naïve representationalist theories in cognitive science that propose such things as "pictures in the head" or that (like good-old-fashioned-AI) treat accurate representations of external properties as general-purpose goal states for cognition. This is all to the good. Yet by dispensing with implementation-level concepts such as predictive inference, it struggles with the important question of what exactly is going on in our heads during the exercise of mastery of a sensorimotor contingency. [9]

4.3 Predictive perception of sensorimotor contingencies

A powerful response is given by integrating SMC theory with PP, in the guise of PPSMC (Predictive Perception of SensoriMotor Contingencies; Seth 2014b). An extensive development of PPSMC is given elsewhere (see Seth 2014b plus commentaries and response). Here I summarize the main points. First, recall that under PP prediction errors can be minimized either by updating perceptual predictions or by performing actions, where actions are generated through the resolution of proprioceptive prediction errors. Also recall that PP is inherently hierarchical, so that at some hierarchical level predictive models will encode multimodal and even amodal expectations linking exteroceptive (sensory) and proprioceptive (motor) sensations. These models generate predictions about linked sequences of sensory and proprioceptive (and possibly interoceptive) inputs corresponding to specific actions, with predictions becoming increasingly modality-specific at lower hierarchical levels. These multi-level predictive models can therefore be understood as instantiating the implicit sub-personal knowledge of sensorimotor constructs underlying SMCs and their acquisition. Put simply, hierarchical active inference implies the existence of predictive models encoding information very much like that required by SMC theory.

The next step is to incorporate the notion of *mastery* of SMCs, which, as mentioned, implies an essentially counterfactual kind of implicit knowledge. The simple solution is to augment the predictive models that animate PP with counterfactual probability densities.[10] As introduced earlier (section 4.1), counterfactually-equipped predictive models encode not only the likely causes of current sensory input, but also the likely causes of fictive sensory inputs conditioned on possible but not executed actions. That is, they encode how sensory inputs (and their expected precisions) would change on the basis of a repertoire of possible actions (expressed as proprioceptive predictions), even if those actions are not performed. The counterfactual encoding of expected precision is important here, since it is on this basis that actions can be selected for their likelihood of minimizing the conditional uncertainty associated with a perceptual prediction. There is a mathematical basis for manipulating counterfactual beliefs of this kind, as shown in a recent model where counterfactual PP drives oculomotor control during visual search (Friston 2014; Friston et al. 2012).[11] Here the main point is that counterfactually-rich predictive models supply just what is needed by SMC theory: an answer to the question of what is going on inside our heads during the exercise of mastery of SMCs.

Counterfactual PP makes sense from several perspectives (Seth 2014b). As mentioned above, it provides a neurocognitive operationalisation of the notion of mastery of SMCs that is central to enactive cognitive science. In doing so it dissolves apparent tensions between enactive

9 At a recent symposium of the AISB society that focused on SMC theory, it was stated that "the main question is how to get the brain into view from an enactive/sensorimotor perspective. […] Addressing this question is urgently needed, for there seem to be no accepted alternatives to representational interpretations of the inner processes" (O'Regan & Dagenaar 2014).

10 See Beaton (2013) for a distinct approach to incorporating counterfactual ideas in SMC theory. Beaton's approach remains squarely within the strongly enactivist tradition.

11 There are also some challenges lying in wait here. For instance, it is not immediately clear how important assumptions like the Laplace approximation can generalize to the multimodal probability distributions entailed by counterfactual PP (Otworowska et al. 2014).

cognitive science and approaches grounded in the Bayesian brain, but only at the price of rejecting the strong enactivist's insistence that internal models or representations—of any sort—are unacceptable.[12] PPSMC also provides a solution to the challenge of accounting for perceptual presence within PP. The idea here is that perceptual presence corresponds to the *counterfactual richness* of predictive models. That is, perceptual contents enjoy presence to the extent that the corresponding predictive models encode a rich repertoire of counterfactual relations linking potential actions to their likely sensory consequences.[13] In other words, we experience normal perception as world-revealing precisely because the predictive models underlying perceptual content specify a rich repertoire of counterfactually explicit probability densities encoding the mastery of SMCs.

A good test of PPSMC is whether it can account for cases where normal perceptual presence is lacking. An important example is synaesthesia, where it is widely reported that synaesthetic "concurrents" (e.g., the inexistent colours sometimes perceived along with achromatic grapheme inducers) are not experienced as being part of the world (i.e., synaesthetes generally retain intact reality testing with respect to their concurrent experiences). PPSMC explains this by noticing that predictive models related to synaesthetic concurrents are counterfactually *poor*. The hidden (environmental) causes giving rise to concurrent-related sensory signals do not embed a rich and deep statistical structure for the brain to learn. In particular, there is very little sense in which synaesthetic concurrents depend on active sampling of their hidden causes. According to PPSMC, it is this comparative *counterfactual poverty* that explains why synaesthetic concurrents lack perceptual presence. SMC theory itself struggles to account for this phenomenon—not least because it struggles to account for synaesthesia in the first place (Gray 2003).

12 There is a more dramatic conflict with "radical" versions of enactivism, in which mental processes, and in some cases even their material substrates, are allowed to extend beyond the confines of the skull (Hutto & Myin 2013).

13 Presence may also depend on the hierarchical depth of predictive models inasmuch as this reflects object-related invariances in perception. For further discussion see commentaries and response to (Seth 2014b).

There are some challenges to thinking that perceptual presence uniquely depends on counterfactual richness. One might think that the more familiar one is with an object, the richer the repertoire of counterfactual relations that will be encoded. If so, the more familiar one is with an object, the more it should appear to be real. But *prima facie* it is not clear that familiarity and perceptual presence go hand-in-hand like this.[14] Also, some perceptual experiences (like the experience of a blue sky) can seem highly perceptually present despite engaging an apparently poor repertoire of counterfactual relations linking sensory signals to possible actions. An initial response is to consider that presence might depend not on counterfactual richness *per se*, but on a "normalized" richness based on higher-order expectations of counterfactual richness (which would be low for the blue sky, for instance). These considerations also point to potentially important distinctions between perceived *objecthood* and perceived *presence*, a proper treatment of which moves beyond the scope of the present paper.

5 Active inference

5.1 Counterfactual PP and active inference

Active inference has appeared repeatedly as an important concept throughout this paper. Yet it is more difficult to grasp than the basics of PP, which involve passive predictive inference. This is partly because several senses of active inference can be distinguished, which have not previously been fully elaborated.

In general, active inference can be harnessed to drive action, or to improve perceptual predictions. In the former case, actions emerge from the minimization of proprioceptive prediction errors through engaging classical reflex arcs (Friston et al. 2010). This implies the existence of generative models that predict time-varying flows of proprioceptive inputs (rather than just end-points), and also the transient reduction of expected precision of proprioceptive prediction

14 Thanks to my reviewers for raising this provocative point.

errors, corresponding to sensory attenuation (Brown et al. 2013).

In the latter case, actions are engaged in order to generate new sensory samples, with the aim of minimizing uncertainty in perceptual predictions. This can be achieved in several different ways, as is apparent by analogy with experimental design in scientific hypothesis testing. Actions can be selected that (i) are expected to *confirm* current perceptual hypotheses (Friston et al. 2012); (ii) are expected to *disconfirm* such hypotheses; or (iii) are expected to *disambiguate* between competing hypotheses (Bongard et al. 2006). A scientist may perform different experiments when attempting to find evidence against a current hypothesis than when trying to decide between different hypotheses. In just the same way, active inference may prescribe different sampling actions for these different objectives.

These distinctions underline that active inference *implies* counterfactual PP. In order for a brain to select those actions most likely to confirm, disconfirm, or decide between current predictive model(s), it is necessary to encode expected sensory inputs and precisions related to potential (but not executed) actions. This is evident in the example of oculomotor control described earlier (Friston et al. 2012). Here, saccades are guided on the basis of the expected precision of sensory prediction errors so as to minimize the uncertainty in current perceptual predictions. Note that this study retained the higher-order prior that only a single perceptual prediction exists at any one time, precluding active inference in its disambiguatory sense.

Several related ideas arise in connection with these new readings of active inference. Seeking disconfirmatory or disruptive evidence is closely related to maximizing Bayesian surprise (Itti & Baldi 2009). This also reminds us that the best statistical models are usually those that successfully account for the most variance with the fewest degrees of freedom (model parameters), not just those that result in low residual error *per se*. In addition, disambiguating competing hypotheses moves from Bayesian model selection and optimization to model comparison, where arbitration among competing models is mediated by trade-offs between accuracy and model complexity (Rosa et al. 2012).

The information-seeking (or "infotropic"[15]) role of active inference puts a different gloss on the free energy principle, which had been interpreted simply as minimization of prediction error. Rather, now the idea is that systems best ensure their long-run survival by inducing the *most predictive* model of the causes of sensory signals, and this requires disruptive and/or disambiguating active inference, in order to always put the current-best model to the test. This view helps dissolve worries about the so-called "dark room problem" (Friston et al. 2012), in which prediction error is minimized by predicting something simple (e.g., the absence of visual input) and then trivially confirming this prediction (e.g., by closing one's eyes).[16] Previous responses to this challenge have appealed to the idea of higher-order priors that are incompatible with trivial minimization of lower-level prediction errors: closing one's eyes (or staying put in a dark room) is not expected to lead to homeostatic integrity on average and over time (Friston et al. 2012; Hohwy 2013). It is perhaps more elegant to consider that disruptive and disambiguatory active inferences imply exploratory sampling actions, independent of any higher-order priors about the dynamics of sensory signals *per se*. Further work is needed to see how cost functions reflecting infotropic active inference can be explicitly incorporated into PP and the free energy principle.

5.2 Active interoceptive inference and counterfactual PP

What can be said about counterfactual PP and active inference when applied to *interoception*? Is there a sense in which predictive models underlying emotion and mood encode counterfactual associations linking fictive interoceptive signals (and their likely causes) to autonomic or allostatic controls? And if so, what phenomeno-

15 Chris Thornton came up with this term (personal communication).

16 The term "dark room problem" comes from the idea that a free-energy-minimizing (or surprise-avoiding) agent could minimize prediction error just by finding an environment that lacks sensory stimulation (a "dark room") and staying there.

logical dimensions of affective experience depend on these associations? While these remain open questions, we can at least sketch the territory.

We have seen that active inference in exteroception *implies* counterfactual processing, so that actions can be chosen according to their predicted effects in terms of (dis)confirming or disambiguating sensory predictions. The same argument applies to interoception. For active interoceptive inference to effectively disambiguate predictive models, or (dis)confirm interoceptive predictions, predictive models must be equipped with counterfactual associations relating to the likely effects of autonomic or (at higher hierarchical levels) allostatic controls. At least in this sense, interoceptive inference then also involves counterfactual expectations.

That said, there are likely to be substantial differences in how counterfactual active inference plays out in interoceptive settings. For instance, it may not be adaptive (in the long run) for organisms to continually attempt to disconfirm current interoceptive predictions, assuming these are compatible with homeostatic integrity. To put it colloquially, we do not want to drive our essential variables continually close to viability limits, just to check whether they are always capable of returning. This recalls our earlier point (section 4.1) that predictive control is more naturally applicable to interoception than exteroception, given the imperative of maintaining the homeostasis of essential variables. In addition, the causal structure of counterfactual associations encoded by interoceptive predictive models is undoubtedly very different than in cases like vision. These differences may speak to the substantial phenomenological differences in the kind of perceptual presence associated with these distinct conscious contents (Seth et al. 2011).

6 Conclusion

This paper has surveyed predictive processing (PP) from the unusual viewpoint of cybernetic origins in active homeostatic control (Ashby 1952; Conant & Ashby 1970). This shifts the perspective from perceptual inference as furnishing representations of the external world for the consumption of general-purpose cognitive mechanisms, towards model-based predictive control as a primary survival imperative from which perception, action, and cognition ensue. This view is aligned with the free energy principle (Friston 2010); however it attempts to account for specific cognitive and phenomenological properties, rather than for adaptive systems in general. Several implications follow from these considerations. Emotion becomes a process of active interoceptive inference (Seth 2013)—a process that also recruits autonomic regulation and influences intuitive decision-making through behavioural allostasis. A common predictive principle underlying interoception and exteroception also provides an integrative view of the neurocognitive mechanisms underlying embodied selfhood, in particular the experience of body ownership (Apps & Tsakiris 2014; Limanowski & Blankenburg 2013; Suzuki et al. 2013). In this view, the experience of embodied selfhood is specified by the brain's "best guess" of those signals most likely to be "me" across exteroceptive and interoceptive domains. From the perspective of cybernetics the embodied self is both that which needs to be homeostatically maintained and also the medium through which allostatic interactions are expressed.

A second influential line deriving from cybernetics sets PP within the broader context of model-based versus enactivist perspectives on cognitive science. On one hand, cybernetics has been cited in support of non-representational cognitive science in virtue of its showing how simple mechanisms can give rise to complex and apparently goal-directed behaviour by capitalizing on agent-environment interactions, mediated by the body (Pfeifer & Scheier 1999). On the other, the cybernetic legacy shows how PP can put mechanistic flesh on the philosophical bones of enactivism, but only by embracing a finessed form of representationalism (Seth 2014b). A key concept within enactive cognitive science is that of mastery of sensorimotor contingencies (SMCs). This concept is useful for understanding the qualitative character of distinct perceptual modalities, yet as expressed within enactivism it lacks a firm implementation basis. "Pre-

dictive Perception of SensoriMotor Contingencies" (PPSMC) addresses this challenge by proposing that SMCs are implemented by predictive models of sensorimotor relations, underpinned by the continuity between perception and action entailed by active inference. *Mastery* of sensorimotor contingencies depends on predictive models of counterfactual probability densities that specify the likely causes of sensory signals that *would* occur *were* specific actions taken. By relating PP to key concepts in enactivism, this theory is able to account for phenomenological features well treated by the latter, such as the experience of perceptual presence (and its absence in cases like synaesthesia).

Considering these issues leads to distinct readings of active inference, which at its most general implies the selective sampling of sensory signals to minimize uncertainty about perceptual predictions. At a finer grain, active inference can involve performing actions to confirm current predictions, to disconfirm current predictions, or to disambiguate competing predictions. These different senses rest on the concept of counterfactually-equipped predictive models; and they generalize the free energy principle to include Bayesian-model comparison as well as optimization and inference.

In summary, the ideas outlined in this paper provide a distinctive integration of predictive processing, cybernetics, and enactivism. This rich blend dissolves apparent tensions between internalist and enactivist (model-based and model-free) views on the neural mechanisms underlying perception, cognition, and action; it elaborates common predictive mechanisms underlying perception and control of self and world; it provides a new view of emotion as active interoceptive inference, and it shows how "counterfactual" predictive processing can account for the phenomenology of conscious presence and its absence in specific situations. It also finesses the concept of active inference to engage distinct forms of hypothesis testing that prescribe different sampling actions (one bonus is that the "dark room problem" is elegantly solved). At the same time, new and difficult challenges arise in validating these ideas experimentally and in distinguishing them from alternative explanations that do not rely on internally-realised inferential mechanisms.

Acknowledgements

I am grateful to the Dr. Mortimer and Theresa Sackler Foundation, which supports the work of the Sackler Centre for Consciousness Science. This work was also supported by ERC FP7 grant CEEDs (FP7-ICT-2009-5, 258749). Many thanks to Thomas Metzinger and Jennifer Windt for inviting me to make this contribution, and for the insightful and helpful reviewer comments they solicited. I'm also grateful to Kevin O'Regan and Jan Dagenaar for inviting me to speak at a symposium entitled "Consciousness without inner models?" (London, April 2014), which provided a feisty forum for debating some of the ideas presented here.

References

Adams, R. A., Shipp, S. & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, *218* (3), 611-643. 10.1007/s00429-012-0475-5

Apps, M. A. & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, *41*, 85-97. 10.1016/j.neubiorev.2013.01.029

Ashby, W. R. (1952). *Design for a brain.* London, UK: Chapman and Hall.

——— (1956). *An introduction to cybernetics.* London, UK: Chapman and Hall.

Aspell, J. E., Heydrich, L., Marillier, G., Lavanchy, T., Herbelin, B. & Blanke, O. (2013). Turning the body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception. *Psychological Science*, *24* (12), 2445-2453. 10.1177/0956797613498395

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76* (4), 695-711. 10.1016/j.neuron.2012.10.038

Beaton, M. (2013). Phenomenology and embodied action. *Constructivist Foundations*, *8* (3), 298-313.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11* (4), 209-243. 10.1177/1059712303114001

Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, *13* (1), 7-13. 10.1016/j.tics.2008.10.003

Bongard, J. (2013). Evolutionary robotics. *Communications of the ACM*, *56* (8), 74-85. 10.1145/2493883

Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, *314* (5802), 1118-1121. 10.1126/science.1133687

Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391* (6669), 756-756. 10.1038/35784

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Brooks, R. A. (1991). Intelligence without reason. In J. Mylopoulos & R. Reiter (Eds.) *Proceedings of the 12th international joint conference on artificial intelligence - volume 1* (pp. 569-595). San Francisco, CA: Morgan Kaufmann Publishers.

Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14* (4), 411-427. 10.1007/s10339-013-0571-3

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there. Putting brain, body, and world together again.* Cambridge, MA: MIT Press.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavior and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

——— (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.

Clark, A. & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58* (1), 7-19. 10.1093/analys/58.1.7

Conant, R. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1* (2), 89-97.

Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13* (4), 500-505. 10.1016/S0959

——— (2009). How do you feel now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10* (1), 59-70. 10.1038/nrn2555

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A. & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7* (2), 189-195. 10.1038/nn1176

Critchley, H. D. & Seth, A. K. (2012). Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron*, *74* (3), 423-426. 10.1016/j.neuron.2012.04.012

Damasio, A. (1994). *Descartes' error.* London, UK: Mac Millan.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M. & Dalgleish, T. (2010). Listening to your heart. How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21* (12), 1835-1844. 10.1177/0956797610389191

Dupuy, J.-P. (2009). *On the origins of cognitive science: The mechanization of mind.* Cambridge, MA: MIT Press.

Evrard, H. C., Forro, T. & Logothetis, N. K. (2012). Von economo neurons in the anterior insula of the macaque monkey. *Neuron*, *74* (3), 482-489. 10.1016/j.neuron.2012.03.003

Ferri, F., Chiarelli, A. M., Merla, A., Gallese, V. & Costantini, M. (2013). The body beyond the body: Expectation of a sensory event is enough to induce ownership over a fake hand. *Proceedings of the Royal Society B: Biological Sciences*, *280* (1765), 20131140-20131140. 10.1098/rspb.2013.1140

Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10* (1), 48-58. 10.1038/nrn2536

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360* (1456), 815-836. 10.1098/rstb.2005.1622

——— (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13* (7), 293-301. 10.1016/j.tics.2009.04.005.

——— (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11* (2), 127-138. 10.1038/nrn2787

——— (2014). Active inference and agency. *Cognitive Neuroscience*, *5* (2), 119-121. 10.1080/17588928.2014.905517

Friston, K. J., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology - Paris*, *100* (1-3), 70-87. 10.1016/j.jphysparis.2006.10.001

Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102* (3), 227-260. 10.1007/s00422-010-0364-z

Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3* (151), 1-20. 10.3389/fpsyg.2012.00151

Friston, K. J., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3* (130), 1-7. 10.3389/fpsyg.2012.00130

Gallese, V. & Sinigaglia, C. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences*, *15* (11), 512-519. 10.1016/j.tics.2011.09.003

Gibson, J. J. (1979). *The ecological approach to visual perception.* Hillsdale, NJ: Lawrence Erlbaum.

Gray, J. A. (2003). How are qualia coupled to functions? *Trends in Cognitive Sciences*, *7* (5), 192-194. 10.1016/S1364-6613(03)00077-9

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *290* (1038), 181-197. 10.1098/rstb.1980.0090

Gu, X., Hof, P. R., Friston, K. J. & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, *521* (15), 3371-3388. 10.1002/cne.23368

Gu, X. & Fitzgerald, T. H. (2014). Interoceptive inference: Homeostasis and decision-making. *Trends in Cognitive Sciences*, *18* (6), 269-270. 10.1016/j.tics.2014.02.001

Hinton, G. E. & Dayan, P. (1996). Varieties of Helmholtz Machine. *Neural Networks*, *9* (8), 1385-1403. 10.1016/S0893

Hohwy, J. (2013). *The predictive mind.* Oxford, UK: Oxford University Press.

——— (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-22). Frankfurt a.M., GER: MIND Group.

Hutto, D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content.* Cambridge, MA: MIT Press.

Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49* (10), 1295-1306. 10.1016/j.visres.2008.09.007

James, W. (1894). The physical basis of emotion. *Psychological Review*, *1*, 516-529.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27* (12), 712-719. 10.1016/j.tins.2004.10.007

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A,Optics, image science and vision*, *20* (7), 1434-1448. 10.1364/JOSAA.20.001434

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neurosciences*, *7* (547), 1-20. 10.3389/fnhum.2013.00547

Makin, T. R., Holmes, N. P. & Ehrsson, H. H. (2008). On the other hand: Dummy hands and peripersonal space. *Behavioural Brain Research*, *191* (1), 1-10. 10.1016/j.bbr.2008.02.041

Metzinger, T. (2003). *Being no one.* Cambridge, MA: MIT Press.

Moseley, G. L., Olthof, N., Venema, A., Don, S., Wijers, M., Gallace, A. & Spence, C. (2008). Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. *Proceedings of the National Academy of Sciences of the United States of America*, *105* (35), 13169-13173. 10.1073/pnas.0803768105

Neal, R. M. & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.) *Learning in Graphical Models* (pp. 355-368). Dordrecht, NL: Kluwer Academic Publishers.

Noë, A. (2004). *Action in perception.* Cambridge, MA: MIT Press.

——— (2006). *Experience without the head.* Clarendon, NY: Oxford University Press.

O'Regan, J. K. & Dagenaar, J. (2014). Consciousness without inner models: A sensorimotor account of what is going on in our heads. *Proceedings of the AISB.* http://doc.gold.ac.uk/aisb50/

O'Regan, J. K., Myin, E. & Noë, A. (2005). Skill, corporality and alerting capacity in an account of sensory consciousness. *Progress in Brain Research*, *150*, 55-68. 10.1016/S0079-6123(05)50005-0

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24* (5), 939-1031.

Otworowska, M., Kwisthout, J. & van Rooj, I. (2014). Counterfactual mathematics of counterfactual predictive models. *Frontiers in psychology: Consciousness Research*, *5* (801), 1-2. 10.3389/fpsyg.2014.00801

Paulus, M. P. & Stein, M. B. (2006). An insular view of anxiety. *Biological psychiatry*, *60* (4), 383-387. 10.1016/j.biopsych.2006.03.042

Pfeifer, R. & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

Pickering, A. (2010). *The cybernetic brain: Sketches of another future.* Chicago, IL: University of Chicago Press.

Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M. & Rawlins, J. N. (1999). Dissociating pain from its anticipation in the human brain. *Science*, *284* (5422), 1979-1981. 10.1126/science.284.5422.1979

Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16* (9), 1170-1178. 10.1038/nn.3495

Quattrocki, E. & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavioral Reviews*, *47C*, 410-430. 10.1016/j.neubiorev.2014.09.012

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2* (1), 79-87. 10.1038/4580

Rosa, M. J., Friston, K. J. & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, *208* (1), 66-78. 10.1016/j.jneumeth.2012.04.013

Salomon, R., Lim, M., Pfeiffer, C., Gassert, R. & Blanke, O. (2013). Full body illusion is associated with widespread skin temperature reduction. *Frontiers in Behavioral Neuroscience*, *7* (65), 1-11. 10.3389/fnbeh.2013.00065

Schachter, S. & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*, 379-399. 10.1037/h0046234

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17* (11), 565-573. 10.1016/j.tics.2013.09.007

——— (2014a). Interoceptive inference: From decision-making to organism integrity. *Trends in Cognitive Sciences*, *18* (6), 270-271. 10.1016/j.tics.2014.03.006

——— (2014b). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience*, *5* (2), 97-118. 10.1080/17588928.2013.877880.

Seth, A. K. & Critchley, H. D. (2013). Interoceptive predictive coding: A new view of emotion? *Behavioral and Brain Sciences*, *36* (3), 227-228.

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2* (395), 1-16. 10.3389/fpsyg.2011.00395

Singer, T., Critchley, H. D. & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13* (8), 334-340. 10.1016/j.tics.2009.05.001

Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R. & Phelps, E. A. (2014). Interoceptive ability predicts aversion to losses. *Cognition and Emotion*, 1-7. 10.1080/02699931.2014.925426

Suzuki, K., Garfinkel, S. N., Critchley, H. D. & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, *51* (13), 2909-2917. 10.1016/j.neuropsychologia.2013.08.014

Thompson, E. & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5* (10), 418-425. 10.1016/S1364-6613(00)01750-2

Tsakiris, M., Tajadura-Jimenez, A. & Costantini, M. (2011). Just a heartbeat away from one's body: Interoceptive sensitivity predicts malleability of body-representations. *Proceedings. Biological sciences / The Royal Society*, *278* (1717), 2470-2476. 10.1098/rspb.2010.2547

Ueda, K., Okamoto, Y., Okada, G., Yamashita, H., Hori, T. & Yamawaki, S. (2003). Brain activity during expectancy of emotional stimuli: An fMRI study. *NeuroReport*, *14* (1), 51-55. 10.1097/01.wnr.0000050712.17082.1c

Van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy*, *92* (7), 345-381. 10.2307/2941061

Varela, F., Thompson, E. & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Verschure, P. F., Voegtlin, T. & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, *425* (6958), 620-624. 10.1038/nature02024nature02024

Wolpert, D. M. & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3 Suppl*, 1212-1217. 10.1038/81497