What might interoceptive inference reveal about consciousness?

Abstract: The mainstream science of consciousness offers a few predominant views of how the brain gives rise to awareness. Chief among these are the Higher-Order-Thought Theory, Global Neuronal Workspace Theory, Integrated Information Theory, and hybrids thereof. In parallel, rapid development in predictive processing approaches have begun to outline concrete mechanisms by which interoceptive inference shapes selfhood, affect, and exteroceptive perception. Here, we consider these new approaches in terms of what they might offer our empirical, phenomenological, and philosophical understanding of consciousness and its neurobiological roots.

Keywords: Interoceptive Inference, Predictive Processing, Consciousness, Self-Inference, Global Neuronal Workspace, Higher Order Thought Theory, Active Inference

Table of Contents

Introduction

What is Consciousness?

If you have ever been under general anaesthesia, you surely remember the experience of waking up. However, this awakening is different from the kind we do every morning, in that it is preceded by a complete lack of subjective experience, a dark nothingness, without even the awareness of time passing. This transition presents a clear insight into the two extremes of conscious experience.

While these strong contrasts delimit the borders of consciousness, you might also consider the phenomenological properties which reveal themselves upon further reflection. Foremost is the unique "mineness" of any conscious experience. In the transition from sleep to wakefulness, there seem to be distinct properties of ownership and agency. Whereas the infinite void of sleep belongs to no one, even before opening my eyes there is a distinct sense in which experience is happening to someone. In phenomenological terms, we can think about this as the minimal, pre-reflexive conditions about which my experiences are uniquely my own (Gallagher, 2000). Consciousness then is something which happens to a sentient subject, which is lived through as the embodied point of view of those seemingly ineffable subjective properties.

A sufficient theory of consciousness then, will deal with each of these properties in turn. What distinguishes conscious states from non-conscious ones? How does selfhood and agency influence these properties? Which sorts of mechanisms give rise to both the phenomenological contents of consciousness, and determine which sorts of states become accessible to conscious thought? How might the body, or emotions, interact with these properties of consciousness?

Answering these questions is no easy task. Certainly, most who study consciousness have heard the joke that there are as many theories of consciousness as there are consciousness theorists. Our goal here is not to provide a comprehensive predictive processing or active inference theory of consciousness, of which there are already a rapidly growing number (for reviews, see Hohwy & Seth, 2020; A. K. Seth & Hohwy, 2020; Whyte, 2019; Whyte & Smith, 2021). Rather, we aim to illustrate how the notion of interoceptive inference and related concepts might inform the theoretical and empirical science of consciousness, by generating alternative process theories that can then be subject to empirical evaluation.

Current mainstream approaches to consciousness can be largely divided into several camps, though the boundaries are fuzzy and hybrid theories abound. Writ large, these include the Global Neuronal Workspace Theory (GNWS), Higher-Order Thought Theory (HOTT), and the Integrated Information Theory (ITT). These theories share some key properties, but also differ substantially in terms of the types of phenomena they seek to explain and the mechanisms they appeal to in doing so. In what follows, we will discuss some of the more obvious places in which predictive processing and interoceptive inference theories tie in with these approaches. Here, we summarize key concepts from some of the leading theories of consciousness and discuss how interoceptive inference might fit into them and inform future theoretical and empirical directions. Our main goals here are the following; first, to accurately and concisely review several of the most

popular theories of consciousness, namely HOTT, GWS, IIT and active inference accounts. We then aim to describe the emerging concept of interoceptive inference, and finally we explore the potential of interoceptive inference to integrate with each of the theories, and how it might illuminate future research directions.

What is Interoceptive Inference?

First, however, we must introduce the standard set pieces of predictive processing and interoceptive inference. Predictive processing can be described as a set of theories which aim to understand how expectations – both neural and psychological – shape, constrain, and ultimately define the mind. These theories have deep roots in cybernetics, information processing, and seminal prospective control models emerging from early 1960s motor and activity theory. A key feature of predictive processing is the basic notion that biological information processing occurs primarily via the minimization of (information theoretic) surprise, such that the nervous system can be understood as a hierarchy of top-down predictions and bottom-up prediction errors. Whilst most early theories extrapolated this basic scheme to explain restricted phenomena such as prospective motor control and the sense of agency (Sperry, 1950; Synofzik et al., 2008; von Helmholtz, 1925), in recent years these approaches have exploded with a myriad of conceptual, computational, and empirical work.

An in-depth review of the scope of predictive processing is beyond this current article. For the unfamiliar reader, we here recall the basic principles, but for a more thorough treatment numerous recent reviews exist, both of the general computational and theoretical principles (Bastos et al., 2012; Clark, 2013; K. Friston, 2009, 2018b; Hohwy, 2013), and their relationship with notions of embodiment and selfhood (Petzschner et al., 2021; A. Seth & Critchley, 2013; A. K. Seth & Friston, 2016).

In summary, these approaches surmise that the brain, much like a Russian Matryoshka or nesting doll, comprises an interlocking hierarchical web, with each unit or level of this web predicting the output of the lower level. At the outermost layer of this hierarchical 'brain web' one finds the sensory epithelium and motor apparatus of the agent – that is, the means by which the agent takes in information about the world external to itself, and acts upon those sensory inputs to alter the world. As one moves from these outermost layers, venturing deeper into the nervous system, neuronal populations encode or invert a model of its inputs[1]. This generative model comprises three key components: a prediction (e.g., of a hierarchically lower expectation), a prediction error (e.g., encoding the difference between the expectation and its prediction), and the precision of each of these signals (e.g., encoding their predictability). This simple motif is replicated from the lowest, most basic neural representations of first order neurons predicting the

[1] invert here is using the technical (Bayesian sense) it refers to the inverse mapping between consequences and causes afforded by a generative model where causes generate consequences. In short, inverting a generative model means inferring the (hidden) causes of (observable) consequences.

activity of sensory effectors, to the highest order, most polymodal representations encoding concepts, selfhood, and preferences.

Early predictive processing theories largely appealed to this motif of prediction error minimization (PEM) to explain phenomena such as visual perception (Rao & Ballard, 1999), motor control (K. Friston, 2011), agency (Synofzik et al., 2008), or social cognitive meta-representation (Kilner et al., 2007; Koster-Hale & Saxe, 2013; Tamir & Thornton, 2018). In contrast, the new "radical predictive processing" wave embraces the unifying nature of the predictive brain in an attempt to explain how all aspects of information processing and behaviour emerge from the integrated hierarchical flow of predictions, prediction errors and their precision (Allen & Friston, 2018; Clark, 2015). Within this framework then, we can consider both the specific hierarchical processing of interoceptive sensations (Allen, 2020; A. K. Seth, 2013a), and the broader implications of embodied, affective inference with respect to our understanding of consciousness.

Interoception is generally used to refer to the sensation, perception, and metacognition of the visceral cycles which govern an agent's homeostasis, allostasis, and ultimately its survival (Barrett & Simmons, 2015; Sherrington, 1952; Vaitl, 1996). This includes, on the ascending side, the sensory information conveying heartbeats, respiration, and the activity of the stomach and gut to the brain – literally, gut feelings. On the descending side, interoception denotes the visceromotor signals and allostatic reflex arcs by which agents maintain their homeostasis in the face of environmental challenges. Interoceptive processes are thus those which enable an agent to monitor and control the bodily states that are necessary to maintain the balance between energy expenditure and consumption.

We can further demarcate interoceptive processes into those which directly subserve *homeostasis*, that is the maintenance of a steady state defined by specific metabolic set-points, and *allostasis*, the proactive control of the body – and environment – to resolve homeostatic needs before they arise (Barrett et al., 2016; Kleckner et al., 2017; Sterling & Eyer, 1988). For example, biological necessity dictates that body temperature, blood oxygenation, and blood glucose level are all maintained within a restrictive range of values. Any sustained deviation from these values is likely to negatively impact an organism's survival, whether through the direct inducement of cellular death, or by the slow attrition of metabolic surplus through starvation. If oxygen is too low, or temperature too high, the brain can directly engage adaptive physiological reflexes, maintaining homeostasis by increasing respiratory frequency or decreasing systolic blood pressure.

These simple sensory-motor reflex arcs, illustrated in Figure 1, can be readily understood by appeal to predictive mechanisms not unlike that of a common household thermostat. That is to say, a low-level spinal, thalamic, or brainstem circuit is generally sufficient to encode the set-point as a prior expectation on the heart-rate, respiratory frequency, or blood pressure. As in afferent control theory, this problem reduces to one of increasing or decreasing the descending visceromotor predictions to minimize any sensory prediction error that occurs: c.f., the equilibrium point hypothesis in motor control (A. G. Feldman, 2009) and related perceptual control theories (Mansell, 2011). One can thus easily envision simple predictive engrams, which monitor visceral

inputs and adjust bodily states as needed to maintain the overall integrity of the system. By comparing the re-afferent sensory inputs to the expected change induced by each top-down prediction, the system can meet whatever thermoregulatory, metabolic, or other homeostatic demands are needed, with relatively little need for higher order cognition.

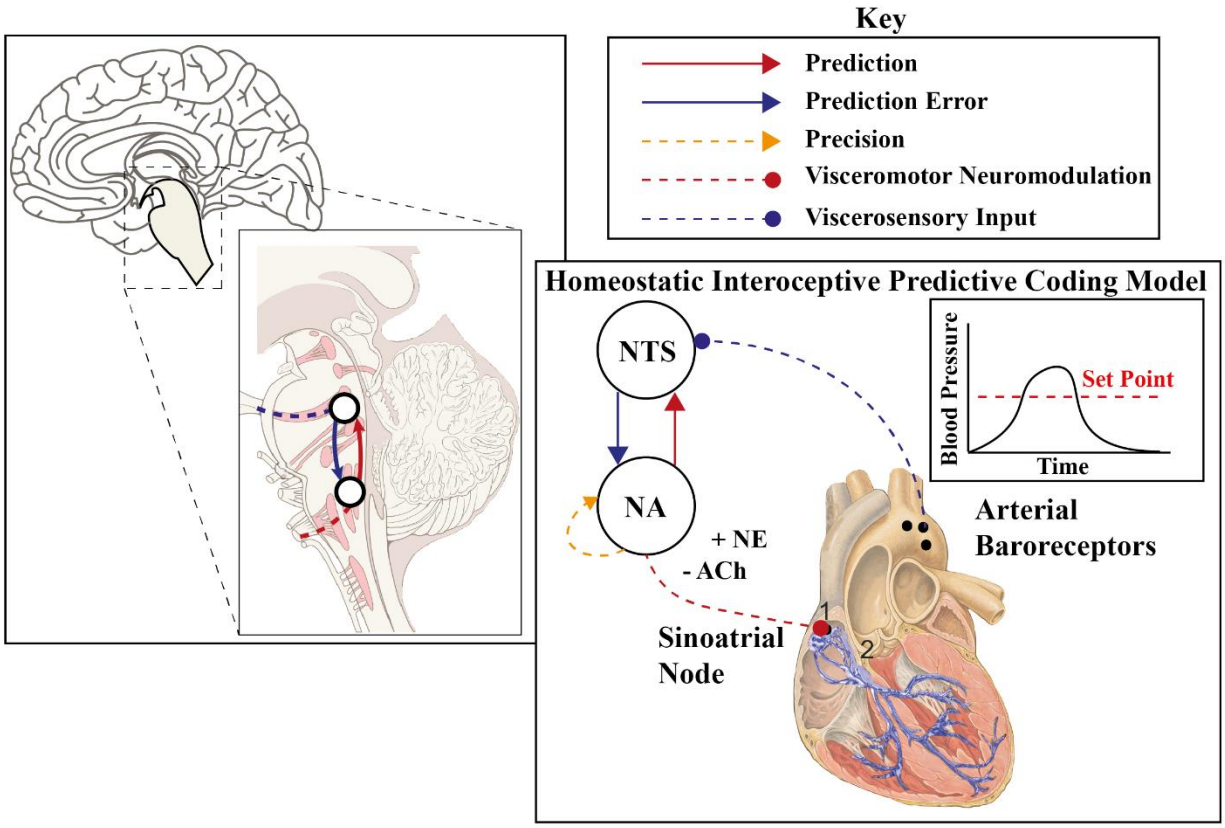

Figure 1: Simplified Homeostatic Control via Interoceptive Inference. This simplified schematic illustrates an example of low-level interoceptive predictive coding in the cardiac domain. Here, a simplified two-node control loop maintains a homeostatic set-point by minimizing the error between afferent cardiac sensory inputs and descending neuromodulatory efferent control. In this example, blood pressure and heart-rate are controlled by a cardiac comparator circuit circumscribed in the primary medulla of the brainstem. Arterial baroreceptors located in the aorta and carotid artery increase their firing rate whenever blood pressure rises above a homeostatic set-point. This firing is relayed via the cranial nerves to the nucleus tractus solitarius (NTS). The NTS acts as a comparator, computing the difference between descending blood pressure predictions and these incoming signals. The difference, or prediction error, is relayed upwards to the nucleus ambiguus (NA), which regulates heart rate via descending cholinergic neuromodulation, triggering the sinoatrial node (SA) to reduce cardiac frequency. An efferent copy (i.e., a descending cardiac prediction) is sent downwards to the NTS, and the comparative loop continues until blood pressure falls below the homeostatic set-point. The relative strength of top-down and bottom-up signals (i.e., their precision) is regulated via neuromodulatory gain control, depicted as self-connections in orange. For illustration purposes, the underlying cardiac neurophysiology has been simplified, leaving out for example the perfused excitatory effects of noradrenaline.

In contrast, allostatic processes are needed whenever the environment or body can no longer maintain these set-points through simple, internal reflex actions alone. For example, if I consistently fail to meet my energy needs, the body will begin to consume itself. Here, merely maintaining homeostasis is insufficient for survival – the agent must identify the external, hidden causes which are causing the increased allostatic load. For example, the environment may no longer contain sufficient resources, in which case the agent should deploy exploratory cognitive mechanisms to find greener pastures. Similarly, if an environment becomes overly threatening (i.e., if the long-term volatility of threats increases), merely increasing or decreasing the heart-rate is no longer sufficient. Instead, I should engage more complex fight or flight routines, to remove the immediate threats and make the situation more amenable to my survival.

Interoceptive inference in the context of allostasis can thus be viewed as operating at a level once (or thrice) removed from that of basic homeostasis. Whereas interoceptive inference at the first order might merely involve the regulation of viscerosensory and visceromotor prediction errors, allostatic interoceptive inference requires the agent to link these low-level variables to contextual ones operating at fundamentally longer timescales. As such, interoceptive inference at this level naturally links to the representation of selfhood, valence, and other metacognitive concepts linking the agent's current homeostatic state to the overall volatility of its environment and conspecifics (Barrett et al., 2016; Petzschner et al., 2017, 2021).

At a broader level still, we can consider the phylogenetic and ontogenetic role that interoceptive processes play in the overall structure and organization of the nervous system. One standout example of this is found in the Free Energy Principle (FEP), a normative biological theory which posits specific foundational and information theoretic constraints on specific biological process theories (K. Friston, 2009). The FEP emphasizes that at the very basis of any biological agent is the self-organized maintenance of its own existence (K. Friston, 2013). In this sense, the very structure of the nervous system can be seen as entailing a generative model[2], which ensures the agent will engage in both homeostatic and allostatic processes. Under the FEP then, the body (both visceral and somato-morphic) are understood as a kind of "first prior" (Allen & Tsakiris, 2018; Ciaunica et al., 2021), which shapes the evolutionary refinement of the predictive mind. Through this lens, the interoceptive hierarchy plays a special role not only in maintaining an agents' survival, but in determining the salience of every action and ensuing belief updating, and ultimately value itself is understood as whatever maximizes the evidence for the agents' model of a survivable world (c.f., "the self-evidencing brain") (Hohwy, 2016).

What then can these set pieces about the brain tell us about consciousness? To start, any predictive processing theory will obviously posit a central role for expectations and predictions in the genesis and contents of consciousness. If the mind is primarily concerned with the representation of future events (i.e., the consequences of action), then it seems likely that

[2] 'Entail' is used carefully here to acknowledge that the generative model is a mathematical construct, not something that is physically realized: neuronal processes can be understood as minimizing free energy that is a function of a generative model; however, neuronal dynamics that are realized reflect *free energy gradients* (that can be cast as a prediction error), not the free energy *per se*.

consciousness is also predominantly prospective. But should a theory of consciousness posit that specific, higher-order modules generate our subjective experience, or rather that it emerges from the collective prediction error minimizing activity of the organism? Similarly, it can be assumed that most predictive processing theories of consciousness should posit a central role in the encoding and modulation of precision – in determining which particular predictions become conscious, and in terms of how conscious predictions should influence affective, metacognitive, and self-related phenomena. That is to say, a basic predictive processing theory of consciousness is likely to ascribe some facets of both *access* and *phenomenal* consciousness (Block, 1995) to error minimizing predictions, and the precision of signals which ensure one particular hypothesis dictates the contents of consciousness versus another.

Must we appeal at all to interoceptive processes in a theory of consciousness? As we shall see, this depends largely on the overarching theory of consciousness developed, i.e., which conscious phenomena are the target of explanation. Certainly a general PEM-based theory of consciousness would ascribe our bodily self-consciousness to the hierarchical minimization of homeostatic and allostatic prediction errors (Ainley et al., 2016). E.g., my consciousness of my heart rate or respiration could be argued to be a product of the viscerosensory and visceromotor prediction errors and precision signals which drive Bayesian belief updating. In this sense, such interoceptive sensations are likely to dominate my awareness, whenever these systems give off prediction errors, whose precision may need updating[3]. Yet such a theory would not posit anything particularly unique about interoceptive inference, casting it as just another parcel of the hierarchical organism which gives rise to various bodily aspects of consciousness. Alternatively, one could develop an FEP or similarly radical predictive-processing theory of consciousness, wherein interoceptive inference may fundamentally underpin access and/or phenomenal consciousness. To consider these different possibilities, we now review predominant theories of consciousness in light of interoceptive inference.

Predictive Higher Order Thought Theory (PHOTT)

Higher order-thought theories (HOTT) stem originally from the analytic philosophy of mind (Carruthers, 2007; Carruthers & Gennaro, 2020; Rosenthal, 2006), yet have also found substantive purchase in the empirical science of consciousness (Brown et al., 2019; H. Lau & Rosenthal, 2011; LeDoux & Brown, 2017). In essence, HOT theories argue that properties of conscious experience arise from the relationship between mental states and higher-order representations of these states (Rosenthal 2005). Critically, this implies that a first-order representation by itself is not part of conscious content, unless it is accompanied by another (higher-order) process that is reflecting on

[3] The updating of the precision of prediction errors is generally read as sensory attention or attenuation Brown, H., R. A. Adams, I. Parees, M. Edwards and K. Friston (2013). "Active inference, sensory attenuation and illusions." Cogn Process **14**(4): 411-427. This speaks to an intimate link between conscious (interception-pointing) inference and attentional selection – or sensory attenuation.

its content. In this sense, HOTT stipulates that an agent can be conscious of some representation X *if and only if* the agent possesses a higher order meta-representation of X. This approach is based on strong assumptions about the links between phenomenal and access consciousness: according to HOTT, conscious states are by definition those that the agent is aware of.

Empirically speaking, HOTT is often associated with metacognitive approaches to modelling consciousness, such as the popular signal-detection theoretic (SDT) framework (Brown et al., 2019; Fleming & Lau, 2014; Ko & Lau, 2012; H. C. Lau, 2007; H. Lau & Rosenthal, 2011). Here, for a conscious state to be labeled as such, an experimental subject must not only exhibit above chance accuracy for detecting some stimulus, but also show explicit conscious awareness of their own accuracy, typically measured via subjective confidence or awareness ratings. Now the metric of consciousness is not just whether a subject can reliably discriminate or detect some input, but whether the subject possesses an accurate meta-representation of their own sensory process (i.e., there should also be a high correlation of confidence and accuracy). Neurobiologically, HOTT proponents frequently argue that the prefrontal cortex plays a crucial role in this metacognitive re-representation of first-order perceptual contents, and as such is sometimes said to be a necessary and sufficient neural correlate of consciousness (NCC).

What then might a "predictive higher-order thought theory" (PHOTT) look like? To our knowledge no theorists have yet directly developed a PHOTT, and a full derivation is beyond the scope (and expertise) of the present article. However, we here briefly sketch some constitutive components of a potential PHOTT, in the hopes of illuminating how interoceptive inference might contribute to such a theory, and in guiding future theoretical, empirical, and computational work.

While thus far no explicit PHOTT theory of consciousness has been proposed, the close alignment of these approaches to empirical and computational metacognition research provides some clear starting points. Metacognition, i.e., the meta-representation and control of first order cognitive or perceptual processes, is typically viewed through a decision-theoretic framework in which the agent must monitor the signal and noise distributions underlying first-order perceptual performance, in order to arrive at a representation of the overall probability that one is making correct responses. Fleming (2020) suggests starting with metacognitive reports of awareness; after all, we can only be aware of another's conscious state through their reports, through language. In the Higher Order State Space (HOSS) model, awareness corresponds to inference on the generative model of the perceptual content, and can be represented as an additional hierarchical state that signals whether perceptual content is present or absent in lower levels. In this case, the higher order thought is cast as a posterior belief over the lower-order contents of consciousness.

Several theorists have proposed Bayesian or predictive extensions of these basic models (Fleming & Daw, 2016; H. C. Lau, 2007; Yeung & Summerfield, 2012), where typically the second-order model is seen as integrating either the contents or the precision of lower-order representations (e.g., the confidence associated with a prediction error encoding a visual input) with high-level "self-priors" describing one's efficacy or overall ability within that cognitive domain. Thus, a basic Bayesian view of metacognition (and meta-representation more generally)

posits an extended cognitive hierarchy in which low-level precision signals are read-out and integrated according to some higher-order self-model.

This raises some immediate set-pieces and questions for a PHOTT model of consciousness. In the philosophical literature the exact nature of the meta-representation needed to render a first order representation conscious has been the subject of intense debate. For example, opposing philosophical camps argue that a HOT must be conceptual in nature to render phenomenal consciousness, versus "higher order perception" (HOP) theorists who posit a kind of "inner sense" theory, which maintains that HOTs need not be conceptual in nature.

Returning to the predictive brain, we find multiple possible candidates for HOTs or HOPs, depending on what particular process theory one works within. For example, in more modular or comparator-based approaches to predictive processing, one could posit the existence of an explicit metacognition module which monitors first-order perceptual representations in order to form an explicit, conceptual HOT encoding the probability that these are correct (as opposed to illusory) percepts. In this sense, predictive higher-order-thoughts (PHOTS) would be ascribed to the higher-order, content-based predictions originating from deep within the brain's hierarchy, encoding relational properties between conscious contents (e.g., the connection between the sensory features encoding a lover's face and the warm affective association therein), or as in the Bayesian metacognitive modules described before, simply encoding the prior probability that a percept is correct given some conceptual self-knowledge and the ongoing pattern of lower-order perceptual prediction errors.

Alternatively, one could argue for a PHOTT (or perhaps a PHOPT) in which the contents of first-order prediction are largely irrelevant to whether a percept becomes conscious or not, and instead emphasize that PHOTs are fundamentally concerned with meta-representing the *precision* of lower-order contents. This aligns both with extant Bayesian theories of metacognition, which emphasize that subjective awareness arises from a posterior estimate of precision, and with the intuitive notion that precision is itself fundamentally a second-order statistic (that is, a meta-representation) of first-order predictive processes. In this case, a precision-focused PHOT would likely emphasize the role of higher-order neural modules in extracting and re-representing the precision (but not the contents) of lower-order predictions, and conscious states would be those associated with the greatest a posteriori precision.

Clearly, these examples are meant to serve as high level outlines illustrating how the set-pieces and explanatory concepts present in predictive processing can be circumscribed within a HOTT of consciousness. Much work remains to be done extrapolating from these basic ideas to a rigorous overall theory. We anticipate that along the way, difficult questions will need to be addressed, concerning for example whether PHOTS are fundamentally concerned with contentful meta-representation, or only with representing the confidence or predictability of first-order processes. One interesting question which emerges immediately, for example, is whether any precision signal could be seen as a sufficient higher-order meta-representation, or whether only higher-order *expected precision* signals would qualify. What we mean is that, according to radical

predictive processing theories (Clark, 2015), precision signals can be found at all levels of the central nervous system (Allen & Tsakiris, 2018; Bruineberg & Rietveld, 2014).

At each level of the brain's canonical microcircuitry then, there is a kind of meta-representation encoding the precision of prediction errors arising at that level, and these local precision signals govern the overall flow of contents through the cortical hierarchy. Are these low-level meta-representations sufficient for a content to become conscious? If so, it would appear then that a PHOT theory of consciousness may help to unify recurrent neural processing and HOTT approaches (Lamme, 2006; Lamme & Roelfsema, 2000), as phenomenal consciousness would emerge from the interaction of local recurrent connections and their associated precision weighting low-level perceptual circuits. In contrast, if it is the explicit representation of expected precision (i.e., top-down, typically poly-modal predictions of future changes in lower order precision) that renders a lower state conscious or not, then the resulting PHOT would likely ascribe neuromodulatory circuits and prefrontal modules as fundamental for determining consciousness (Boly et al., 2017; Odegaard et al., 2017).

How does interoception fit into the PHOTT framework? One option is that interoceptive information, just like visual input, is another source of lower order perceptual input, which can be integrated with other information and reflected upon by higher order processes to become a subject of conscious experience. In this sense then, PHOTs predicting either higher-order interoceptive contents (e.g., the association between multiple viscerosensory systems and affect or value) would largely determine whether one is conscious or not of any given interoceptive sensation. In this sense, interoception would not play any special role in a PHOT theory of consciousness, other than offering another channel of perceptual contents which may be configured within any other higher order thoughts or percepts.

Alternatively, if the preferred PHOTT emphasizes the role of meta-representations encoding expected precision, then interoceptive processes may play a more constitutive role in determining either phenomenal or access consciousness. Generally speaking, the optimization of expected precision has been proffered as a unifying mechanism by which salience, attention, and high-level self-control emerge (H. Feldman & Friston, 2010; Kanai et al., 2015; Moran et al., 2013; Parr & Friston, 2017, 2019). Furthermore, the very capacity to supply low levels of hierarchical inference with predictions of precision or predictability has been proposed as a necessary condition for qualitative experience; in the sense of precluding phenomenal transparency (Limanowski, 2017; Limanowski & Blankenburg, 2013; Limanowski & Friston, 2018).

This approach views bottom-up and top-down attention as emergent properties of minimizing "precision-prediction errors", such that the top-down control of expected precision can selectively enhance or inhibit lower-order percepts. Interoceptive prediction errors and precision thereof are here thought to play a unique role in determining what is salient for an agent in any given context, such that unexpected challenges to homeostasis or allostasis essentially govern the innate value of different outcomes. Computational and conceptual models have expanded on this view to describe a process of metacognitive and interoceptive self-inference, in which the a priori expected precision afforded the homeostatic and allostatic fluctuations is always higher than that of over

sensory-motor channels (Allen, 2020; Allen et al., 2019, 2020). As fluctuations in, for example, blood temperature or arterial pulsation, can directly modulate the noise (i.e., inverse precision) of neuronal circuits in a global fashion (Chow et al., 2020), then the representation of expected precision is argued to both sample directly from the precision of interoceptive prediction errors, and to utilize descending visceromotor control as a means of optimizing sensory precision.

In PHOTT terms then, this could be taken as an argument that visceral prediction and precision signals play an especially important role in the meta-representation of first-order perceptual contents, such that either their subjective salience is largely governed by higher-order thoughts about the interaction between the visceral body and the exteroceptive sensorium. In this sense then, both the "shape" or "contours" of phenomenal consciousness, and the likelihood that a percept becomes conscious (i.e., access consciousness) may depend in part on the top-down meta-representation of expected (interoceptive) precision. Such a process theory would then show some alignment between the PHOTT approach, and recent theoretical proposals suggesting that interoceptive signals may play a fundamental role in shaping the "mineness" or subjective quality of conscious experiences (Ainley et al., 2012, 2016; Azzalini et al., 2019; Fotopoulou & Tsakiris, 2017; A. K. Seth & Tsakiris, 2018).

GNWS and Interoceptive Inference

The Global Workspace theories (Baars, 1988; Baars & Franklin, 2007; Dehaene & Changeux, 2011) originate from the idea that consciousness arises from processing of information in the brain, and the way in which specific information is selected and broadcast across the brain in order to generate a coherent representation. Here, the brain is composed of a set of specialized, local cortical processing units, which are richly interconnected by excitatory pyramidal neurons spanning between frontal and parietal regions. A piece of information, represented in one or several of the processing units, can cross a threshold and be selected for broadcasting (i.e., amplification) in the process of 'ignition', whereby it is simultaneously made available to all processing units. For example, when a bird perches nearby and chirps, my attention is drawn to the sound, and I will gaze around to find the source. The perceptual inputs associated with the bird are carried up and processed, and as they enter the global workspace and become 'globally available' as a part of consciousness, such that, they, along with the idea of the bird and the feeling the moment is associated with in my body, are broadcast to various brain systems.

These may include memory allowing me to remember the moment, motor action, or higher cognitive systems which enable me to make decisions and talk about my experience. Crucially, most information that is available to and processed by the brain need not enter the global workspace, here consciousness is about how and which information is selected for global processing and awareness. The Global Neuronal Workspace (GNW) theory (Dehaene et al., 2006; Dehaene & Changeux, 2011) specifies that information which does become available to the global workspace (GW) then recruits brain networks extending over frontal and parietal regions which can integrate the dispersed sources of information into a coherent conscious phenomenon. Thus,

the prefrontal cortex plays a central role within GNW as in HOT theories, yet they differ in what functions they ascribe to it (Mashour et al., 2020); in HOT the higher-order metacognitive processes representing first-order states are what constitute consciousness, so if they are in the PFC, this region becomes a source of consciousness. In GNW meanwhile, conscious states emerge by the broadcasting of information across systems, which can happen due to long range connections between PFC, other fronto-parietal regions comprising the GW. We emphasize that HOTT and GNW are not mutually exclusive, and in fact, several works aim to bridge and unify these theories (Dehaene et al., 2017; Graziano et al., 2019). The Attentional Schema Theory for example, merges GNW and HOT by proposing that attention amplifies signals so that they may reach ignition, and that there is a higher order representation of the GW which represents the dynamics and implications of having a GW, which is what gives rise to phenomenological consciousness.

Unifying approaches are also building active inference-based models within the GNW framework. The predictive global neuronal workspace (PGNW) (Hohwy, 2013; Whyte, 2019; Whyte & Smith, 2021) combines Bayesian active inference with experimentally corroborated components of the GNW. The PGNW enables us to examine one of the core questions arising from the GNW theories: what determines the ignition threshold? Within predictive processing, the information that crosses the threshold to reach ignition is that which best accommodates PEs throughout the hierarchy, so that the best-fitting (PEM) model of the world is selected and broadcast across systems (Friston, Breakspear et al. 2012). Ignition then represents the point at which an evidence accumulation process has reached the threshold where it becomes the most likely explanation of the world (i.e., the current ascending input). The PGNW therefore represents ignition as an inferential process. As in the active inference framework below, ignition here requires sufficient temporal thickness to coordinate and contextualize lower levels of processing (Whyte & Smith, 2021). In order to be able to speak of my experience of the chirping bird, I need a representation that is maintained for some period of time and that extends back in time to include me observing the bird.

According to the standard GNW account the anterior insula, a key hub processing visceral information and involved in interoceptive awareness (Critchley et al., 2004; Evrard, 2019), selects and prioritizes information prior to possible amplification by the GW (Michel, 2017). Another theory in the same spirit (Chanes & Barrett, 2016), presents the limbic cortex (including the anterior insula, anterior cingulate cortex, among other areas) as the ‘limbic workspace’ in light of the rich bi-directional connections between these areas and lower levels of processing. In this view, cortical lamination is a distinguishing feature, so that predictions move up from less to more laminated areas while PEs move down in the opposite direction.

Within the PGNW view, interoception is a perceptual system (or set of systems), sensing the internal states and rhythms occurring in the body, and information from it can independently or together with congruent information from other systems, be broadcast by the GNW. For example, I may become aware of a sudden stomach cramp, which incites me to think about what I have eaten earlier in the day. However, recent evidence proposes that interoception might also play a

modulating role on other systems (Chanes & Barrett, 2016; Michel, 2017), whereby interoceptive prediction errors and/or precisions affect the likelihood that other modules are brought into the GW, driving ignition itself through the modulation of salience.

It has been suggested that the brain maintains a self-model representing the status of the body, which is continuously updated to fit ascending interoceptive input by changing interoceptive PEs (Barrett & Simmons, 2015). Generally, in these accounts, what achieves ignition can be understood as the relationship between the expected (top-down) and sensory (bottom-up) precision, where when the self-model increases the precision of lower order modules, they become better fitting models of the world and are more likely to reach ignition. Further work has proposed that interoception may play a crucial role within the self-model, by either conditioning expected precision (Allen et al., 2019, 2020), or by modulating the degree to which lower-order representations are interpreted as related to the sense of self (Apps & Tsakiris, 2014; Azzalini et al., 2019; Babo-Rebelo et al., 2016; Limanowski & Blankenburg, 2013; Quigley et al., 2021; A. K. Seth, 2013b).

Thus, as we saw in the previous section, depending on the exact predictive process theory one motivates, interoception may act simply as one of many modules within the GNWS, or it may play a more foundational role, either by guiding the top-down selection of modules into the WS by the self-model, or by enhancing the gain or precision associated with lower-order, non-interoceptive modules as to alter their probability of promotion into the WS. We therefor propose that future experimental and computational work will likely benefit from modelling how interoceptive processes interact with conscious processing of stimuli, and the proposed neurophysiological signatures of ignition, such as the P300 component, to ultimately understand whether interoceptive prediction errors or their precision alter the process of ignition and the overall topology of the GNWS.

IIT and Interoceptive Inference

The Integrated Information Theory (IIT) (Oizumi et al., 2014; Tononi et al., 2016) of consciousness is an attempt at a formal method for mathematically describing the conscious experience of any given system, agnostic as to all but the causal structure of its substrate. The theory focuses on making an intrinsic description of the system, that is, how the system is to itself, opposed to an extrinsic description from the perspective of an outside observer. The IIT takes as starting point five axioms for what constitutes any phenomenological experience. From that, five criteria are derived which must be met in order for a physical system to support conscious experience: 1) the system must exist, that is, exert and be subject to causal power; and it must do so over itself in a way that is 2) structured of component elements; 3) informative i.e., distinguishable from other causal states; 4) integrated or unitary as a whole, and irreducible to independent subsets; and 5) exclusive or definite, specifying its own borders.

To measure the degree to which a system fulfils these criteria, a measure of *integrated conceptual information* is used, denoted as Φ, which measures the degree to which the system

exerts causal power over itself in a way that is irreducible to the activity of its components. The conscious parts of a system are called complexes and are those parts of the system that specify the highest Φ without overlapping with one other. The axiom that complexes cannot overlap also means that smaller complexes are not conscious, even if Φ is larger than zero, as long as they are contained within a larger complex with a higher Φ. Conversely, a large complex is also not conscious if there are smaller complexes within it with a higher Φ. This leads to predictions of consciousness in the brain being situated in areas with more integrated connections, currently thought to be an temporo-parietal-occipital hot-zone in the posterior cortex (Koch et al., 2016). This excludes more feed-forward networks like the cerebellum, explaining why this structure does not obviously contribute to consciousness despite its large number of neurons (Tononi & Koch, 2015). The exclusivity axiom also means that experience only happens at one spatial and temporal scale of organization, namely at the level at which Φ is highest (Hoel et al., 2016).

The conscious experience of a complex involves concepts, which are causal mechanisms within the complex that specify irreducible cause-effect repertoires. All the concepts together form a concept structure, which can be interpreted as a geometric shape in a multidimensional concept space. The concept structure of a complex is thought to reflect the *content* of consciousness, while the size of Φ reflects the *amount* of consciousness as a whole. Importantly, the concept structure not only depends on the current state of the complex, but also on the other possible states it could take, since it is defined by how the system causally constraints itself. This allows for the possibility for negative concepts, that is, the absence of some state (e.g., not-red), and that conscious experience is also enriched by the increase of more possible states (e.g., seeing green includes not-red, not-blue, etc.).

The IIT and predictive processing theories of brain function and of consciousness take quite different starting points in a range of respects: IIT is concerned with understanding how systems in general relate to themselves, while predictive processing addresses how the brain, specifically, relates functionally to the surrounding environment, including the body. The former begins entirely in describing phenomenology to identify compatible types of physical systems, while the latter largely takes the opposite direction and starts with what is required for the physical brain in order to self-organize and maintain itself, going from there to describe phenomenology. This makes it challenging to combine the two approaches, a project that is far beyond the scope of this paper – but see (K. J. Friston et al., 2020) for a discussion, in terms of the information geometry of active inference.

It might be worth briefly speculating, however, what predictive processing accounts might be able to offer IIT to inform the broader discussion of interoceptive processing and consciousness. One notion is that hierarchical, precision-weighted prediction error belief updating schemes might provide (neuronal) structures that result in high levels of integration, a suggestion that might potentially be investigated by calculating Φ of canonical neuronal schemas from predictive processing and comparing it to other proposed schemas. The prediction error minimization loops in PEM theories are certainly more complex and integrated than the zero Φ feedforward networks in, say, artificial neural networks. Zooming out, one might also ask if the overall structure of the

brain relates to the level of integration; does, for example, the presence of a self-model in the brain somehow constitute or allow for higher levels of integration? One could certainly imagine that the part of the brain that constitutes the 'self-as-hypothesis' might be highly integrated, given that it has to coordinate impressions from many brain areas – and that can be parsimoniously explained as being caused and sampled by 'self-as-agent'. In that case, one might expect high integration in the deepest (highest) parts of the predictive hierarchy; i.e., instantiated in interactions between the default mode network or the salience network (which we note, are also key hubs for interoceptive processing), where the self-model might be instantiated (Margulies et al., 2016). It is also possible that the presence of a higher order integrative component of the larger network is not, in itself, sufficiently integrated to constitute the conscious part of the brain – but that its presence and monitoring allows other parts to be integrated enough to become conscious. The monitoring of the self-model essentially underwrites homeostasis, that is, self-maintenance, which must be tightly related to the exertion of causal power over or the causal constraining of oneself. Indeed, it has been argued by Marshall and colleagues (2017) that intrinsic control and maintenance of causal borders is characteristic of living systems, which seems to align with active inference and predictive coding formulations.

Further, IIT's image of components of a system forming concepts, that in turn can form higher order concepts through integration – for example, integrating the single notes of a song into a melody – might suit such a thing as a self-model particularly well, for the self-model might be thought of as the highest order concept integrating all those lower-order concepts that relate to oneself. This should in particular integrate concepts somehow related to the body, such as interoceptive processes, with perceptual concepts about the current environment as well as those about the agent itself. Finally, it is worth noting that the fact that a higher amount of negative concepts result in higher levels of Φ suits well the argument that counterfactual depth is related to consciousness (Corcoran et al., 2020). Having negative concepts at least conceptually (if not formally) seems related to having a model or experience of the world that describes not only what is, but also what could have been, providing one platform where the otherwise very different theories might meet.

Now, how might interoceptive inference fit into IIT's story of consciousness? Initially, the fit seems poor here as well; for IIT is concerned with the consciousness of systems in general, and additionally also mainly concerned with the experience of these systems *independently* of the external world around them. Interoceptive inference is mainly defined specifically in relation to the brain making inferences about the body within which it is located. We must therefore first allow our conceptualization of interoception to cover any conscious system's inferences about any kind of body – be it that of a human, animal, plant or complex machine. In addition, we must assume that the conscious experience of a system under IIT must have some kind of relation (structural, perhaps, rather than representational) to the surrounding environment, including the body. This assumption should be treated with caution however, as bridging phenomenological and more functional accounts in this way is no simple project. Here, we offer a speculative outline of these potential links for further discussion.

For example, from the view of IIT, one can define the body, generally, as a part of the environment that situates conscious processing, and that it must both react to and control in order to persist, as well as to navigate the rest of the environment. In this view, homeostasis simply becomes acting on or controlling directly that part of the environment that is always present and that I am tightly coupled to, the body, while allostasis is recast as acting on the rest of the environment through the body. It should then be likely that any complex system has in its concept structure some concepts related to bodily states. These concepts need not be about the body *per se*; they can be experienced in any way, as long as they are a result of the system navigating within and controlling its body. This means that emotions, understood as embodied-inference (Barrett, 2017; Hesp, Smith, et al., 2019), can certainly act as concepts within the system's concept structure that are not in themselves experienced as part of the body, but rather as part of experience itself. One might also hypothesize that conscious systems with complex bodies that need complex behaviour to control and navigate their environment must also be more integrated, and have a richer concept structure that allows for a diverse variety of emotions – in line with how systems become more conscious if they evolve to navigate in a more complex external environment (Albantakis et al., 2014). In this way, a more complex body could directly afford a richer experience with more options and nuances for emotive concepts and the concurrent higher number of negative concepts: a hypothesis that could in principle be investigated in simulation studies. In particular, it may be that high demands on and capabilities for allostasis require a system to be highly integrated and result in a rich bodily experience, since homeostasis by itself is arguably simple.

In IIT, conscious experience occurs when a system is able to constrain its own future in a way irreducible to its component elements. Interoceptive inference, then, is inferences about the survival probabilities of the system itself or at least its nearest and most intimate surroundings, the body. Interoceptive inference is therefore crucial for a system to be able to self-constrain in very complex environments. The brain certainly depends on it in order to survive, which can be seen as a type of self-constraining. Successful interoceptive inference may also allow the brain to be more integrated with the body; in IIT terms, that is, to couple with the body in an interdependent way. Given that parts of the brain are so highly integrated that their Φ levels are higher than that between brain and body, probably it is unrealistic (even if theoretically possible) that the brain and body would be so integrated that they together would form one conscious complex; but the adaptive value of high integration will still be in effect even if only a part of the brain stays conscious.

One could also imagine that something like a heart – that is, a rhythmical oscillating state which is strongly connected to the rest of the body and brain – would have great effect on a conscious system's concept structure and experience (Allen et al., 2019). It may thus be intriguing to develop evolutionary simulations such as those of Albantakis and colleagues (2014), but with agents that have minimal bodies and task-relevant rhythmically oscillating states that affect the conscious 'brain', too see if such agents evolve concepts relevantly similar to emotions, indicating a phenomenological experience of a bodily state.

Notions of selfhood are also important for some theories of consciousness. What might selfhood be in IIT, and would it be related to interoception? Selfhood in IIT could be thought to be the higher-order concept that integrates all body and self-related concepts (emotions, action possibilities and tendencies, in general all that could be called either homeostasis or allostasis). Because the underlying concepts are integrated into a higher order concept, they are not experienced as separate components, but as an integrated whole, meaning that self is the integration of body-related concepts in the same way that a melody is the integration of the single experienced nodes. This might also suggest a fundamental self-other distinction; for it is possible that the components of the system that underwrite bodily and self-experiences are more integrated with each other than they are with experiences related to the external world. This seems likely given that those components might all be influenced by changes in bodily states and therefore be more co-dependent than changes in the external world. Finally, in another sense, there is a 'self' in IIT in the sense that there is a main complex which is conscious, a centre of consciousness that is arguably separate from homeostatic and allostatic processes. Would there be a concept within the concept structure specifically related to this - or is it rather the entire concept structure, that is, the entirety of experience as an integrated whole, that might here be called the self, and from which the sense of 'mineness' comes? There might here be an opportunity for an I vs. me distinction; i.e., a distinction between the self as the subject and object within experience (Gallagher, 2000). The former being the entirety of integrated consciousness, and the latter being those concepts in my concept structure that are integrated to form a general experience of what I am and can do, expressly based in bodily experiences. Speculatively, the former might correspond to a lower level primary consciousness – sometimes called C1, for example as in (Frith, 2019) – that does not have meta-representations but is still essentially experienced phenomenologically, while the latter might be a form of higher-order consciousness (C2 and higher, and underwriting access consciousness).

Active Inference, Interoception and Consciousness

Active inference is a process theory for how adaptive self-organizing systems come to comply with the normative framework of the Free Energy Principle, and thereby stay in existence (K. Friston et al., 2017). There are several theories of how active inference processes might relate to conscious experience; in the following, we first give a brief introduction to active inference under the free energy principle, and then discuss the existing related consciousness theories. Finally, we consider the potential role of interoception within these approaches and active inference in general.

The Free Energy Principle (K. Friston, 2010, 2019) is a normative principle, essentially stating any self-organizing system that maintains a non-equilibrium steady state must, in order to resist random perturbations and maintain itself, act as if it minimized its variational free energy, or maximized the Bayesian model evidence, of its implicit model of the world, given sensory observations. This is often situated in an across-scales blanket-oriented formal ontology where reality is described as a nested hierarchy of Markov Blanket structures, that is, statistical

separations of internal states from external states (Clark, 2017; Hesp, Ramstead, et al., 2019; Kirchhoff et al., 2018).

Blanket states are separated into active states that affect the external world, and sensory states which affect internal states based on impressions from the external world; maintaining a Markov Blanket entails maintaining a non-equilibrium steady state, which mandates gradient flows on variational free energy[4]. These gradient flows mean that, on average, internal states come to statistically model the external world. Furthermore, active states conform to Hamilton's principle of least action so that, on average, active states minimise the path integral of variational free energy over time. Active inference is then a process theory describing *how*, exactly, self-organizing systems might come to minimize their variational free energy now, and in the future (K. Friston et al., 2013; Sajid et al., 2021). On this view, self-organizing systems appear to simulate the consequences of actions in order to select those actions that lead to the least free energy in the future (i.e., least action), leading to a balance between exploratory, information-seeking behaviour, and exploitative, pragmatic behaviour. Active inference (often modelled using Partially Observable Markov Decision Processes) has been used to describe a variety of phenomena, ranging from stratospheric adaption (Rubin et al., 2020) through cellular organization (Kuchling et al., 2020), interoceptive processes (Allen et al., 2019), and neuronal activity (Isomura et al., 2020; Isomura & Friston, 2018) to psychiatric disorders (R. A. Adams et al., 2013; Benrimoh et al., 2018).

Active inference is a formal description of how a system interacts with its environment in order to maintain some desired state, and does not necessarily relate inherently to questions about consciousness. There has, however, been work investigating which types of active inference might underlie conscious experience. Most importantly, it has been argued that consciousness is a result of the generative model implied by the system having temporal and counterfactual depth. That is, that it includes future consequences of actions, and that it includes what would have happened had it acted differently in the past (Corcoran et al., 2020; K. Friston, 2018a). Active inference has also been used to answer the meta-hard problem of consciousness (i.e., why we as researchers are so puzzled by the relation between phenomenal experience and reality). It is also argued that complex agents might come to form mid-level beliefs within their hierarchical models of the world as especially certain, but simultaneously come to realize that these beliefs are irreducibly different from the world. This leads to an inferred chasm between the agent's experiences and the external world, and a seeming irreducible difference between subjective experience and objective reality (Clark et al., 2019). Finally, it is argued that the blanket-oriented ontology described before offers a natural separation between intrinsic information geometries on one side, describing how internal states evolve probabilistically over time, and extrinsic information geometries on the other, describing probabilistic beliefs about external states which are then parametrized by internal states, thus uniting the mind/matter distinction under a monist framework (K. J. Friston et al., 2020).

[4] A gradient flow is simply a description of states that change in the direction of steepest descent on some function of their current value; here, variational free energy.

In addition to this, it is theorized that consciousness is the felt effect that results from explicitly evaluating the expected free energy under different actions, as opposed to automatic or reflexive behaviour (Solms & Friston, 2018). There is also an attempt by Ramstead et al. (Ramstead et al., 2021) to apply generative modelling to understand phenomenology on its own terms, arguing that raw sensory experience can be likened to the observations of an active inference agent, and that the coherent lived experience is then the most likely posterior belief or the best explanation for those raw experiences. Many of these approaches are probably consistent or at least overlap with predictive instantiations of HOT and GNWS theories for consciousness in the brain, for they also emphasize hierarchically structured predictions of the consequences of – and the accuracy of – own beliefs. It should also be noted that there are current attempts at synthesizing consciousness theories like Integrated Information Theory and Global Neuronal Workspace theory under the Free Energy Principle to produce a new Integrated World Modelling Theory of consciousness (Safron, 2019, 2020).

In this section we take pains to distinguish between active inference, as the general process of acting adaptively by making predicted or preferred states most likely through action in a free energy minimizing fashion, and predictive coding, which is a process that commits to a specific kind of message passing in the brain and how it might come to effectuate such active inference. The two can indeed be closely related, as often seen in the literature, e.g. in (R. Adams et al., 2015), but for clarity we keep them separate for now. This also allows us to distinguish between brain-specific consciousness theories relevant for interoception, and those more general statements about consciousness in complex systems that relate to active inference in general. We focus on the latter here; the former brain-specific predictive processing-based approaches to consciousness have been considered in the previous section.

As with the discussion of interoception and IIT above, one can define the body as an external (to the brain or the conscious system, i.e., outside the Markov Blanket) environment that nonetheless is so closely coupled with the brain that it follows it around everywhere, making the body at once both the most important part of the external environment to monitor and predict on one side, and to control on the other. From here it is not a stretch to claim that there can, indeed, in general, not be any successful active inference without at least a rudimentary kind of interoception, for active inference rests on predicting the consequences of one's actions upon the world (and thereby on one's own sensory observations); since the body realizes this influence of actions on the world, then a failure to properly model and make inferences about the body also leads to a (fatal) failure to affect the world in an autopoietic way.

This means that successful self-maintaining systems must always model their bodies, be they humans, plants or machines, and that the structure of the body and the actions it can effectuate, therefore, should be strongly determinant for the types of experiences an organism has. One might also consider that the system-within-the-body might model itself *as* the body, that is, in order to reduce unnecessary complexity of its generative model simply coarse-grain itself and its body into a whole in the self-model. This, of course, is only feasible (i.e., free-energy minimizing) if the body and the controlling system (for example the brain) are so tightly coupled that distinguishing

between them has only irrelevant advantages to the agent's resulting behaviour. In addition, given that maintenance of the body is crucial for the controlling system's self-maintenance, the pragmatic value – that is, prior preferences over outcomes – for an active inference agent should largely be defined in terms of – or at least in relation to – bodily states, and therefore interoception, largely in accordance with the idea of homeostatic priors and inference as having a privileged position as a first prior (Allen et al., 2020; Allen & Tsakiris, 2018). This clearly posits interoception, self-maintenance, emotion, interoceptive inference, value and consciousness as tightly interlinked concepts.

When approaching consciousness and interoception from the perspective of active inference of self-organizing systems in general, rather than specifically from predictive processing in the brain, one might ask why one should focus particularly on the boundary between the brain on the one side, and the body and external world on the other? Markov Blanket partitions can be constructed on many levels of neuronal organization, from single neurons to brain regions (K. J. Friston et al., 2021; Hipólito et al., 2021), indicating that active inference occurs on all those levels, at each level potentially displaying either of the qualities associated with consciousness in the discussion above. Focusing on the level of the brain as a whole situated within the body and the external world would be the traditional choice in consciousness research. It is also the level on which bodily processes such as respiration and heartbeats are part of the immediate environment, and therefore where interoceptive inference happens as it is traditionally conceptualized. Indeed, brain inferences about the body and the world are more similar to personal experiences, compared, for example, to inferences made by a brain region about other brain regions. It might also be hypothesized that the level of the brain as a whole is indeed the level of description with, for example, the longest temporal and counterfactual depth, making it the most interesting level for the purposes of consciousness research.

Active inference based accounts of consciousness might be considered functionalistic, because active inference as a framework is centered around how a system interacts with its environment. This is contrasted by Integrated Information Theory (IIT), for example, which focuses on a system's internal causal structure irrespective of its sensorimotor exchanges with the external world. One might imagine two functionally identical (in terms of their blanket states) systems with different internal causal structures, which should therefore have identical generative models from an active inference perspective, but which would have different conscious experiences according to IIT (as in (Oizumi et al., 2014)). A full discussion of the differences between these two approaches are beyond the scope of this paper, but we note that it is a potentially interesting line of research to clarify the theoretical and formal relations between them, for example investigating whether temporal and counterfactual depth of the implied generative model of a system is related to its level of integration.

Active inference formalizations have of course already been brought to bear on the question of interoception, in general: volitional control of respiration can be seen as an active inference process which alters interoceptive models (Boyadzhieva & Kayhan, 2020); interoceptive inference has been related to psychopathologies (Paulus et al., 2019); and it lies at the foundation of theories

of interoceptive inference in general (A. K. Seth, 2013b). Another recent line of work also tries to understand interoception as a type of active self-inference modulating the volatility of sensory-motor representations. We relate this to consciousness in the following, penultimate section.

Interoceptive Self-Inference: an integrated theory of consciousness?

Finally, we consider how the emerging framework of interoceptive self-inference (Allen, 2020; Allen et al., 2020; Allen & Tsakiris, 2018) might offer an integrative approach to the empirical and theoretical study of consciousness. The theory of interoceptive self-inference is a computational and theoretical model which aims to explain how bodily and interoceptive processes shape exteroceptive and metacognitive awareness, and *vice versa*. Interoceptive self-inference can be seen as a process theory built in part from the FEP, based on empirical and phenomenological observations (Allen & Friston, 2018; Gallagher & Allen, 2018). In particular, the theory posits three core observations:

I. To persist, agents must learn to navigate a volatile, ever-changing world (Piray & Daw, 2020a, 2020b; Pulcu & Browning, 2019).
II. Visceral, homeostatic rhythms directly influence the volatility of both lower-order sensory-motor representations (Allen et al., 2019; Chow et al., 2020; Livneh et al., 2020), and metacognitive inferences thereof (Allen et al., 2016; Hauser et al., 2017).
III. Therefore, agents actively infer their own volatility trajectories, in part, by sampling and controlling interoceptive rhythms, resulting in close coupling between top-down expected volatility and the visceral body (Lawson et al., 2021; Petzschner et al., 2017, 2021).

The theory thus proposes that, when estimating our own future reliability or precision, agents intrinsically sample from and predict their own visceral rhythms. Conversely, on shorter timescales, agents can optimize the 'signal-to-noise' ratio of ongoing sensorimotor dynamics through ballistic alterations of those same visceral rhythms (Galvez-Pol et al., 2020; Grund et al., 2021). A simple example here is that of the trained sharpshooter, who modulates their breathing in order to align the timing of a trigger pull with the quiescent period. Interoceptive self-inference is thus the implicit, preconscious or prenoetic process by which the confidence and salience of the sensorium is aligned to the rhythms of the body: we literally self-infer our own precision trajectories, and in doing so, we actively shape them.

Clearly this process of self-inference aligns closely with philosophical and empirical work which describes the importance of an intrinsic predictive self-model, which contextualizes and embodies phenomenal consciousness (Hohwy, 2016; Limanowski & Blankenburg, 2013; T. Metzinger, 2007). Here we further argue that the minimal-self, i.e., the pre-reflective nature of

perceptual consciousness, is closely tied to the interoceptive body, in virtue of the close coupling of these rhythms with the overall stability, reliability, and predictability of the agent's own trajectory. Although the visceral body is rarely the focus of the perceiving self, the interoceptive self-inference model posits that the overall contents of consciousness, and in particular the idiosyncratic salience maps which differ between persons and contexts, are likely to be shaped by the close coupling between expected volatility, sensory-motor precision, and visceral rhythms. Interoceptive self-inference then predicts that sampling of the interoceptive trajectory can be used to estimate the volatility of external states. Cognitive and perceptual biases (e.g., exteroceptive and metacognitive) may then arise from treating interoceptive noise as exteroceptive, such that experimental modulation of interoceptive noise could shift the cognitive bias, as partially demonstrated in recent investigations of interoception and metacognition (Allen et al., 2016; Legrand et al., 2020, 2021) . In parallel, conscious experience may then entail the prioritisation of environmental stimuli which are pertinent to the body's contingencies; for example by increasing the salience of the smell of food when we are hungry.

Is interoceptive self-inference then itself an integrative theory of consciousness? Certainly, in light of the previous discussion, we can find links between PHOTT and PGNWS approaches and self-inference. On the self-inference account, the global workspace itself is cast as a dynamic, prospective self-model, which accumulates evidence from cortical and sub-cortical systems to infer an overall estimate of expected precision. Interoceptive prediction errors are thus cast as a controlling factor in the overall bifurcation, topology, and probability-to-ignition of the global workspace. Speculatively, one could potentially re-describe "ignition" as the process of active inference by which a top-down model is self-inferred, meaning, in which the agent engages neuromodulatory and visceromotor processes to actively reshape or reconfigure the overall landscape or topology of precision, literally bringing the moment-to-moment self into existence. This would imply that "ignition" is itself a process of active self-inference, in which the agent entertains one hypothesis over another regarding the overall shape and functionality of the cortical manifold, maintained through the estimation and control of expected precision.

Similarly, there are clear potential links between PHOTT and interoceptive-self inference. Interoceptive self-inference was originally developed as a model explaining how and why visceral signals impinge upon metacognitive judgements in other, non-interoceptive domains (Allen et al., 2016; Hauser et al., 2017). Metacognition is typically modelled using a signal-detection theoretic approach, in which subjective confidence or awareness is assumed to depend upon a higher-order meta-representation of first-order signal versus noise, plus some additional metacognitive noise (for review, see the earlier section on PHOTT). Interoceptive self-inference inverts this picture, to suggest that metacognitive estimation is a process of self-inferring the probable correlation between the sensorium and ongoing visceral fluctuations. As a silly example, consider the metacognitive evaluation of whether one will do well on an exam: the confidence estimate here depends both on a judgement of expertise within the domain, and perhaps on whether the agent has been binge-drinking the night before and will thus be suffering from sickness behaviours

during the exam. The projection of self-reliability into the future is closely coupled both to domain-relevant knowledge, and the prediction of self-volatility.

Interoceptive self-inference would then align itself somewhere between PHOTT and PGNWS, seeking to explain how and why interoceptive prediction errors and precisions are coupled to the cortical hierarchy to shape both top-down predictions of precision, and to actively infer future self-states through descending visceromotor control. However, we wish to pump the brakes a bit here – PGNWS and PHOTT are both currently under-defined process theories. It remains to be seen whether these or any predictive-processing derived theory of consciousness is empirically productive. That is to say, we believe that the ultimate test of a theory of consciousness should not be whether it neatly ties together different conceptual approaches, but whether it can make clear contrasting predictions regarding the mechanisms underlying consciousness itself. And while interoceptive self-inference does make clear empirical predictions about the linkages between say, learning, metacognition, and interoception, it remains to be seen whether these predictions will be similarly fruitful for consciousness research.

Conclusion

We have reviewed some contemporary approaches to consciousness research in the burgeoning predictive processing literature, with an aim of discovering how research on interoception can inform these emerging discussions. In particular, we highlight links between explanatory concepts found in approaches such as higher-order thought theory, the global neuronal workspace, integrated information theory, and predictive processing versions of these. While our review is by design speculative, we hope to have provided the reader with an overview that can serve as a roadmap for future research in these domains. Overall, we propose that further refinement of the existing theories with consideration for interoceptive inference will prove stimulating to the field. Working out the shared commitments between these different approaches is certainly a monumental endeavour, but one which we hope will ultimately prove fruitful.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, *4*. https://doi.org/10.3389/fpsyt.2013.00047

Adams, R., Friston, K., & Bastos, A. (2015). Active Inference, Predictive Coding and Cortical Architecture. *Recent Advances On The Modular Organization Of The Cortex*, 97–121. https://doi.org/10.1007/978-94-017-9900-3_7

Ainley, V., Apps, M. A. J., Fotopoulou, A., & Tsakiris, M. (2016). 'Bodily precision': A predictive coding account of individual differences in interoceptive accuracy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160003. https://doi.org/10.1098/rstb.2016.0003

Ainley, V., Tajadura-Jiménez, A., Fotopoulou, A., & Tsakiris, M. (2012). Looking into myself: Changes in interoceptive sensitivity during mirror self-observation. *Psychophysiology*, *49*(11), 1672–1676.

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLOS Computational Biology*, *10*(12), e1003966. https://doi.org/10.1371/journal.pcbi.1003966

Allen, M. (2020). Unravelling the Neurobiology of Interoceptive Inference. *Trends in Cognitive Sciences*, *24*(4), 265–266. https://doi.org/10.1016/j.tics.2020.02.002

Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *ELife*, *5*, e18103. PubMed. https://doi.org/10.7554/eLife.18103

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482. https://doi.org/10.1007/s11229-016-1288-5

Allen, M., Legrand, N., Correa, C. M. C., & Fardo, F. (2020). Thinking through prior bodies: Autonomic uncertainty and interoceptive self-inference. *Behavioral and Brain Sciences*, *43*. https://doi.org/10.1017/S0140525X19002899

Allen, M., Levy, A., Parr, T., & Friston, K. J. (2019). In the Body's Eye: The Computational Anatomy of Interoceptive Inference. *BioRxiv*, 603928. https://doi.org/10.1101/603928

Allen, M., & Tsakiris, M. (2018). The body as first prior: Interoceptive predictive processing and the primacy. In *The Interoceptive Mind: From Homeostasis to Awareness* (Vol. 27).

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

Azzalini, D., Rebollo, I., & Tallon-Baudry, C. (2019). Visceral Signals Shape Brain Dynamics and Cognition. *Trends in Cognitive Sciences*, *23*(6), 488–509. https://doi.org/10.1016/j.tics.2019.03.007

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B. J., & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. *Neural Networks: The Official Journal of the International Neural Network Society*, *20*(9), 955–961. https://doi.org/10.1016/j.neunet.2007.09.013

Babo-Rebelo, M., Richter, C. G., & Tallon-Baudry, C. (2016). Neural Responses to Heartbeats in the Default Network Encode the Self in Spontaneous Thoughts. *Journal of Neuroscience*, *36*(30), 7829–7840. https://doi.org/10.1523/JNEUROSCI.0262-16.2016

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23. https://doi.org/10.1093/scan/nsw154

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160011. https://doi.org/10.1098/rstb.2016.0011

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews. Neuroscience*, *16*(7), 419–429. https://doi.org/10.1038/nrn3950

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. https://doi.org/10.1016/j.neuron.2012.10.038

Benrimoh, D., Parr, T., Vincent, P., Adams, R. A., & Friston, K. (2018). Active Inference and Auditory Hallucinations. *Computational Psychiatry (Cambridge, Mass.)*, *2*, 183–204. https://doi.org/10.1162/cpsy_a_00022

Block, N. (1995). On a Confusion About a Function of Consciousness. *Brain and Behavioral Sciences*, *18*(2), 227–247. https://doi.org/10.1017/s0140525x00038188

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *Journal of Neuroscience*, *37*(40), 9603–9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2019.06.009

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00599

Carruthers, P. (2007). Higher-order theories of consciousness. *The Blackwell Companion to Consciousness*, *10*, 9780470751466.

Carruthers, P., & Gennaro, R. (2020). Higher-Order Theories of Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/consciousness-higher/

Chanes, L., & Barrett, L. F. (2016). Redefining the Role of Limbic Areas in Cortical Processing. *Trends in Cognitive Sciences*, *20*(2), 96–106. https://doi.org/10.1016/j.tics.2015.11.005

Chow, B. W., Nuñez, V., Kaplan, L., Granger, A. J., Bistrong, K., Zucker, H. L., Kumar, P., Sabatini, B. L., & Gu, C. (2020). Caveolae in CNS arterioles mediate neurovascular coupling. *Nature*, *579*(7797), 106–110. https://doi.org/10.1038/s41586-020-2026-1

Ciaunica, A., Constant, A., Preissl, H., & Fotopoulou, A. (2021). *The First Prior: From Co-Embodiment to Co-Homeostasis in Early Life*. PsyArXiv. https://doi.org/10.31234/osf.io/twubr

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2015). Radical Predictive Processing. *The Southern Journal of Philosophy*, *53*(S1), 3–27. https://doi.org/10.1111/sjp.12120

Clark, A. (2017). How to Knit Your Own Markov Blanket. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*.

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing Qualia Consciousness as Inference, Not Raw Datum. *Journal of Consciousness Studies, 26*.

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). *From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and the Origins of Cognition*. https://www.preprints.org/manuscript/201911.0083/v2

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

Dehaene, S., & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron, 70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204–211. https://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Evrard, H. C. (2019). The Organization of the Primate Insular Cortex. *Frontiers in Neuroanatomy, 13*. https://doi.org/10.3389/fnana.2019.00043

Feldman, A. G. (2009). New insights into action–perception coupling. *Experimental Brain Research, 194*(1), 39–58.

Feldman, H., & Friston, K. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience, 4*. https://doi.org/10.3389/fnhum.2010.00215

943 Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of*
944 *Consciousness*, *2020*(niz020). https://doi.org/10.1093/nc/niz020
945 Fleming, S. M., & Daw, N. D. (2016). Self-evaluation of decision-making: A general Bayesian
946 framework for metacognitive computation. *Psychological Review*, *124*(1), 91.
947 https://doi.org/10.1037/rev0000045
948 Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human*
949 *Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00443
950 Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of
951 interoceptive inference. *Neuropsychoanalysis*, *19*(1), 3–28.
952 https://doi.org/10.1080/15294145.2017.1294031
953 Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive*
954 *Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005
955 Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews*
956 *Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787
957 Friston, K. (2011). What Is Optimal about Motor Control? *Neuron*, *72*(3), 488–498.
958 https://doi.org/10.1016/j.neuron.2011.10.018
959 Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475.
960 https://doi.org/10.1098/rsif.2013.0475
961 Friston, K. (2018a). Am i self-conscious? (Or does self-organization entail self-consciousness?).
962 *Frontiers in Psychology*, *9*(APR). https://doi.org/10.3389/fpsyg.2018.00579
963 Friston, K. (2018b). Does predictive coding have a future? *Nature Neuroscience*, *21*(8), 1019–
964 1021. https://doi.org/10.1038/s41593-018-0200-7

965 Friston, K. (2019). A free energy principle for a particular physics. *ArXiv:1906.10184 [q-Bio]*.
966 http://arxiv.org/abs/1906.10184
967 Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A.
968 (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1),
969 211–251. https://doi.org/10.1162/netn_a_00175
970 Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the Origins of Consciousness:
971 From Cartesian Duality to Markovian Monism. *Entropy*, *22*(5), 516.
972 https://doi.org/10.3390/e22050516
973 Friston, K., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active
974 Inference, Curiosity and Insight. *Neural Computation*, *29*(10), 2633–2683.
975 https://doi.org/10.1162/neco_a_00999
976 Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013).
977 The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*,
978 *7*. https://doi.org/10.3389/fnhum.2013.00598
979 Frith, C. D. (2019). The neural basis of consciousness. *Psychological Medicine*, 1–13.
980 https://doi.org/10.1017/S0033291719002204
981 Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science.
982 *Trends in Cognitive Sciences*, *4*(1), 14–21. https://doi.org/10.1016/S1364-6613(99)01417-
983 5
984 Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social
985 cognition. *Synthese*, *195*(6), 2627–2648. https://doi.org/10.1007/s11229-016-1269-8

Galvez-Pol, A., McConnell, R., & Kilner, J. M. (2020). Active sampling in visual search is coupled to the cardiac cycle. *Cognition*, *196*, 104149. https://doi.org/10.1016/j.cognition.2019.104149

Graziano, M., Guterstam, A., Bio, B., & Wilterson, A. (2019). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology*, *37*, 1–18. https://doi.org/10.1080/02643294.2019.1670630

Grund, M., Al, E., Pabst, M., Dabbagh, A., Stephani, T., Nierhaus, T., & Villringer, A. (2021). Respiration, heartbeat, and conscious tactile perception. *BioRxiv*, 2021.03.22.436396. https://doi.org/10.1101/2021.03.22.436396

Hauser, T. U., Allen, M., Purg, N., Moutoussis, M., Rees, G., & Dolan, R. J. (2017). Noradrenaline blockade specifically enhances metacognitive performance. *ELife*, *6*. https://doi.org/10.7554/eLife.24901

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. (2019). A Multi-scale View of the Emergent Complexity of Life: A Free-Energy Proposal. In G. Y. Georgiev, J. M. Smart, C. L. Flores Martinez, & M. E. Price (Eds.), *Evolution, Development and Complexity* (pp. 195–227). Springer International Publishing. https://doi.org/10.1007/978-3-030-00075-2_7

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K., & Ramstead, M. (2019). *Deeply Felt Affect: The Emergence of Valence in Deep Active Inference*. https://doi.org/10.31234/osf.io/62pfd

Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience & Biobehavioral Reviews*, *125*, 88–97. https://doi.org/10.1016/j.neubiorev.2021.02.003

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness, 2016*(niw012). https://doi.org/10.1093/nc/niw012

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Hohwy, J. (2016). The Self-Evidencing Brain. *Noûs*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II). https://doi.org/10.33735/phimisci.2020.II.64

Isomura, T., & Friston, K. (2018). In vitro neural networks minimise variational free energy. *Scientific Reports*, *8*(1), 16926. https://doi.org/10.1038/s41598-018-35221-w

Isomura, T., Shimazaki, H., & Friston, K. (2020). Canonical neural networks perform active inference. *BioRxiv*, 2020.12.10.420547. https://doi.org/10.1101/2020.12.10.420547

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169. https://doi.org/10.1098/rstb.2014.0169

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, *15*(138), 20170792. https://doi.org/10.1098/rsif.2017.0792

Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., Quigley, K. S., Dickerson, B. C., & Feldman Barrett, L. (2017). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Human Behaviour*, *1*(5), 0069. https://doi.org/10.1038/s41562-017-0069

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1401–1411. https://doi.org/10.1098/rstb.2011.0380

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, *79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kuchling, F., Friston, K., Georgiev, G., & Levin, M. (2020). Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of Life Reviews*, *33*, 88–108. https://doi.org/10.1016/j.plrev.2019.06.001

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579. https://doi.org/10.1016/S0166-2236(00)01657-X

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in Brain Research* (Vol. 168, pp. 35–48). Elsevier. https://doi.org/10.1016/S0079-6123(07)68004-2

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Lawson, R. P., Bisby, J., Nord, C. L., Burgess, N., & Rees, G. (2021). The Computational, Pharmacological, and Physiological Determinants of Sensory Learning under Uncertainty. *Current Biology, 31*(1), 163-172.e4. https://doi.org/10.1016/j.cub.2020.10.043

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*(10), E2016–E2025. https://doi.org/10.1073/pnas.1619316114

Limanowski, J. (2017). (Dis-)Attending to the Body. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. MIND Group. https://doi.org/10.15502/9783958573192

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00547

Limanowski, J., & Friston, K. (2018). 'Seeing the Dark': Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00643

Livneh, Y., Sugden, A. U., Madara, J. C., Essner, R. A., Flores, V. I., Sugden, L. A., Resch, J. M., Lowell, B. B., & Andermann, M. L. (2020). Estimation of Current and Future Physiological States in Insular Cortex. *Neuron*. https://doi.org/10.1016/j.neuron.2019.12.027

Mansell, W. (2011). Control of perception should be operationalized as a fundamental property of the nervous system. *Topics in Cognitive Science*, *3*(2), 257–261.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, *113*(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113

Marshall, W., Kim, H., Walker, S. I., Tononi, G., & Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2109), 20160358. https://doi.org/10.1098/rsta.2016.0358

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Metzinger, T. (2007). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in Brain Research* (Vol. 168, pp. 215–278). Elsevier. https://doi.org/10.1016/S0079-6123(07)68018-2

Michel, M. (2017). A role for the anterior insular cortex in the global neuronal workspace model of consciousness. *Consciousness and Cognition*, *49*, 333–346. https://doi.org/10.1016/j.concog.2017.02.004

Moran, R. J., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., & Friston, K. J. (2013). Free Energy, Precision and Learning: The Role of Cholinergic Neuromodulation. *Journal of Neuroscience*, *33*(19), 8227–8236. https://doi.org/10.1523/JNEUROSCI.4255-12.2013

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception? *Journal of Neuroscience*, *37*(40), 9593–9602. https://doi.org/10.1523/JNEUROSCI.3217-16.2017

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, *7*(1), 14678. https://doi.org/10.1038/s41598-017-15249-0

Parr, T., & Friston, K. J. (2019). Attention or salience? *Current Opinion in Psychology*, *29*, 1–5.

Paulus, M. P., Feinstein, J. S., & Khalsa, S. S. (2019). An Active Inference Approach to Interoceptive Psychopathology. *Annual Review of Clinical Psychology*, *15*(1), 97–122. https://doi.org/10.1146/annurev-clinpsy-050718-095617

Petzschner, F. H., Garfinkel, S. N., Paulus, M. P., Koch, C., & Khalsa, S. S. (2021). Computational Models of Interoception and Body Regulation. *Trends in Neurosciences*, *44*(1), 63–76. https://doi.org/10.1016/j.tins.2020.09.012

Petzschner, F. H., Weber, L. A. E., Gard, T., & Stephan, K. E. (2017). Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis. *Biological Psychiatry*, *82*(6), 421–430. https://doi.org/10.1016/j.biopsych.2017.05.012

Piray, P., & Daw, N. D. (2020a). A simple model for learning in volatile environments. *PLOS Computational Biology*, *16*(7), e1007963. https://doi.org/10.1371/journal.pcbi.1007963

Piray, P., & Daw, N. D. (2020b). Unpredictability vs. Volatility and the control of learning. *BioRxiv*, 2020.10.05.327007. https://doi.org/10.1101/2020.10.05.327007

Pulcu, E., & Browning, M. (2019). The Misestimation of Uncertainty in Affective Disorders. *Trends in Cognitive Sciences*, *23*(10), 865–875. https://doi.org/10.1016/j.tics.2019.07.007

Quigley, K. S., Kanoski, S., Grill, W. M., Barrett, L. F., & Tsakiris, M. (2021). Functions of Interoception: From Energy Regulation to Experience of the Self. *Trends in Neurosciences*, *44*(1), 29–38. https://doi.org/10.1016/j.tins.2020.09.008

Ramstead, M. J., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G., Smith, R., Dumas, G., Lutz, A., Friston, K., & Constant, A. (2021). *From generative models to generative passages: A computational approach to (neuro)phenomenology*. PsyArXiv. https://doi.org/10.31234/osf.io/k9pbn

Rao, R. P. N., & 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. https://doi.org/10.1038/4580

Rosenthal, D. M. (2006). Consciousness and Higher-Order Thought. In *Encyclopedia of Cognitive Science*. American Cancer Society. https://doi.org/10.1002/0470018860.s00149

Rubin, S., Parr, T., Da Costa, L., & Friston, K. (2020). Future climates: Markov blankets and active inference in the biosphere. *Journal of The Royal Society Interface*, *17*(172), 20200503. https://doi.org/10.1098/rsif.2020.0503

Safron, A. (2019). *Integrated World Modeling Theory (IWMT) Revisited*. PsyArXiv. https://doi.org/10.31234/osf.io/kjngh

Safron, A. (2020). An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and

Characterizing Agentic Causation. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00030

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation*, *33*(3), 674–712. https://doi.org/10.1162/neco_a_01357

Seth, A., & Critchley, H. (2013). Extending predictive processing to the body: Emotion as interoceptive inference. *The Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X12002270

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

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

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160007. https://doi.org/10.1098/rstb.2016.0007

Seth, A. K., & Hohwy, J. (2020). Predictive processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, *0*(0), 1–2. https://doi.org/10.1080/17588928.2020.1838467

Seth, A. K., & Tsakiris, M. (2018). Being a Beast Machine: The Somatic Basis of Selfhood. *Trends in Cognitive Sciences*, *22*(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

Sherrington, C. (1952). *The integrative action of the nervous system*. CUP Archive.

Solms, M., & Friston, K. (2018). How and why consciousness arises: Some considerations from physics and physiology. *Journal of Consciousness Studies*, *25*(5–6), 202–238.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, *43*(6), 482.

Sterling, P., & Eyer, J. (1988). Allostasis: A new paradigm to explain arousal pathology. In *Handbook of life stress, cognition and health* (pp. 629–649). John Wiley & Sons.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239. https://doi.org/10.1016/j.concog.2007.03.010

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

Vaitl, D. (1996). Interoception. *Biological Psychology*, *42*(1), 1–27. https://doi.org/10.1016/0301-0511(95)05144-9

von Helmholtz, H. (1925). Helmholtz's treatise on physiological optics, (Southall JP, transl.). *New York: Optical Society of America*.

Whyte, C. J. (2019). Integrating the global neuronal workspace into the framework of predictive processing: Towards a working hypothesis. *Consciousness and Cognition*, *73*, 102763. https://doi.org/10.1016/j.concog.2019.102763

Whyte, C. J., & Smith, R. (2021). The predictive global neuronal workspace: A formal active inference model of visual consciousness. *Progress in Neurobiology*, *199*, 101918. https://doi.org/10.1016/j.pneurobio.2020.101918

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416