A beautiful loop:
An active inference theory of consciousness

Ruben Laukkonen[1], Karl Friston[6], & Shamil Chandaria[2,3,4,5]

[1] Faculty of Health, Southern Cross University, Gold Coast, Australia
[2] Centre for Eudaimonia and Human Flourishing, Linacre College, Oxford University, UK
[3] Centre for Psychedelic Research, Division of Brain Sciences, Imperial College London, UK
[4] Institute of Philosophy, The School of Advanced Study, University of London, UK
[5] Fitzwilliam College, University of Cambridge, UK
[6] Institute of Neurology, University College London, UK

ABSTRACT

Can active inference model consciousness? We offer three conditions implying that it can. The first condition is the simulation of a reality or generative world model, which determines what can be known or acted upon; namely an *epistemic field*. The second is inferential competition to enter the world model. Only the inferences that coherently reduce long-term uncertainty win, evincing a selection for consciousness that we call *Bayesian binding*. The third is *epistemic depth*, which is the recurrent sharing of the Bayesian beliefs throughout the system. Due to this recursive loop — in a hierarchical system (such as a brain) — the world model contains the knowledge that it exists. This is distinct from self-consciousness, because the world model knows itself non-locally and continuously evidences this knowing (i.e., *field-evidencing*). Formally, we propose a hyper-model for precision-control across the entire hierarchy, whose latent states (or parameters) encode and control the overall structure and weighting rules for all layers of inference. This *Beautiful Loop Theory* is deeply revealing about meditation, psychedelic, and altered states, minimal phenomenal experience, and provides a new vision for conscious artificial intelligence.

Key words: Consciousness; Awareness; Active Inference; Predictive Processing; Free Energy; Meditation; Psychedelics; Sleep; Dreaming; Unconscious; Bayesian Inference; Artificial Intelligence; Neuroscience; Computational Modelling

1. INTRODUCTION

Consciousness is perhaps the biggest mystery in science. At a certain point, most fields of inquiry find that the strange capacity of organisms to experience cannot be overlooked. Books, articles, and media discussing the nature of consciousness abound, with unique perspectives emerging from psychologists, neuroscientists, philosophers, phenomenologists, computer scientists, biologists, physicists, and contemplatives. Yet, most would agree that consciousness remains an inconvenient enigma in a world that otherwise seems reducible to things, objects, patterns, and equations.

Equally, it is clear that the tools of science *can* reveal something about the nature of consciousness. Thousands of experiments attest that consciousness has predictable characteristics, predictable correlates, and fluctuates under predictable conditions (Koch et al., 2016; Frith, 2021). Thanks to this growing evidence base, an array of impressive theories of consciousness (ToCs) have emerged in recent years (Rosenthal, 2000; Seth & Bayne, 2022; Carruthers, 2017; Tononi, 2008; Baars, 2005). These theories have many strengths and explanatory power but a general consensus in the scientific community does not exist. Even the metaphysical assumptions underlying the science of consciousness still result in fierce debates (Kuhn, 2024; Fleming et al., 2023; Kastrup, 2008).

Here, we aim to contribute to these theories by building on a promising general theory of organisms, known as *active inference* or *predictive processing,* under the *free energy principle* (Friston, 2010; Clark, 2013; Hohwy, 2013; Seth & Tsakiris, 2018). Several others have proposed that active inference may provide solutions to different features of conscious experience (e.g., Hohwy, 2022; Hohwy & Seth, 2020; Safron, 2020; 2022; Carhart-Harris et al. 2014; Rudrauf et al. 2017; Friston, 2018; Williford et al. 2018; Clark, 2019; Kanai et al., 2019; Chang et al. 2020; Deane, 2021; Whyte & Smith, 2021; Whyte et al., 2024). Nevertheless, it remains unclear whether active inference can satisfy the conditions for a ToC. Yet others have proposed that we should think of active inference as providing "...theories for consciousness science, rather than ToCs per se" (Seth & Bayne, 2022, p. 446). The question arises, what conditions would active inference need to satisfy in order to cross the ToC threshold? Why is it that the theory is so successful in explaining perception, cognition, and action, but not consciousness itself?

To address these questions, we propose three conditions that seem necessary for consciousness and show how some active inference systems satisfy them. The first condition is a generative world model, or *epistemic field*. This provides the 'space' or contents that can be known, hence the term *epistemic* (Metzinger, 2020). The second condition is *inferential competition*, which determines what becomes conscious and why it is coherent (i.e., addressing the binding problem). The third and final condition is *epistemic depth*, which refers to the fact that the epistemic field is recursively, and widely (i.e., deeply) shared throughout the system. As we will see, this idea has some similar characteristics to "broadcasting" (Dehaene et al., 2003), "information integration" (Tononi, 2008), and "fame in the brain" (Dennett, 2001), albeit with important differences.

Below, we will introduce one condition at a time. We will then show how active inference can provide a parsimonious explanation for a range of cognitive processes and states of consciousness when these conditions are satisfied. Our view also implies that the most basal or minimal form of awareness is a highly simplified (*nearly* contentless) world model knowing itself non-locally. Therefore, a first person perspective, self-modeling, and agency, are not prerequisites of awareness, but are rather local or "contracted" forms of consciousness (Metzinger, 2020).

To be concise, we will avoid an extensive literature review (see Seth & Bayne, 2022; Frith, 2021; or Lau, 2022 for reviews on ToCs). However, as noted above, many features of the theory (reviewed in Table 2 of the discussion) are consistent with elements of other ToCs such as *Global Neuronal Workspace Theory* (GNWT, Dehaene et al., 2003), *Higher Order Theories* (HOT, Lau and Rosenthal, 2011), *Recurrent Processing Theory* (RPT, Lamme and Roelfsema, 2000; Pennartz et al., 2019) and *Integrated Information Theory* (IIT, Tononi, 2008). Links with existing theories will be made throughout. The strength of our approach will be in showing how the interactions of a minimal set of computational assumptions within active inference may provide the ingredients for consciousness, with implications for understanding various states from lucid dreaming to meditation, to psychedelics, as well as artificial intelligence.

2. SIMULATING A REALITY MODEL

In order to move about in a world and keep ourselves alive, we need a model of that world. One could not walk, jump, pick up a glass, catch a ball, or give a hug, without a simulation of the unfolding present. Achieving such a coherent model of reality is an immense feat, especially considering that the brain has to deal with unpredictable, imprecise and often incoherent data (Treisman, 1996). Yet, somehow, our experience of reality seems to make sense—it has depth, color, shape, thought, emotion, people, and objects that we seem to understand and predict with relative ease. Notably, there can also be an awareness of the contents of this world model. We *experience* the catching of the ball, the texture of grass under our feet, and the warm embrace. Our world appears to be alive.

We term this 'experienced world', which is the organism's entire lived reality, the *generative phenomenal, unified* world model (hereafter simply *reality model*). It is *generative* because processes internal to the organism play a central role in constructing or generating the 'output' of the model. It is *phenomenal* because there can be an experience of the reality model—it constitutes our lived world. And it is *unified* because it is coherent or appears to be 'bound' together as a whole. This reality model is also an *epistemic field* because it is a place (or flow of sensations) that can be known, explored, interrogated, and updated—a kind of affordance for action, both physical and mental (Metzinger, 2017). Our model of reality tells us what is possible and what is not possible, what will keep us alive and what will harm us, and even what we ourselves are. We take such a model to be a necessary condition for consciousness because it defines what *can* become conscious, be known, or experienced.

Active inference provides a straightforward solution to — or description of — how organisms construct a reality model apt for their lived world (Friston, 2006; 2010). Various details of this theory will be introduced later; here, it is sufficient to lay out a few of the key tenets. Active inference can be derived from two codependent assumptions: (1) the imperative for biological systems to maintain a boundary between themselves and the environment (i.e., *existence*), and (2) the necessity to remain in a specific set of (characteristic) states compatible with continued existence (i.e., *adaptive actions*). From these premises, we can construct a framework where existence necessitates a generative model of the self and environment, allowing the system to resolve surprising sensations — i.e., *homeostasis* — and anticipate surprising outcomes and maintain themselves through adaptive action, i.e., *allostasis*. In order to learn and update the model, organisms reduce prediction errors, or uncertainty[1]. Or flipped around, the organism persists by seeking evidence for its own existence, i.e., *self-evidencing* (Hohwy, 2016). Minimizing prediction errors—the difference between top-down predictions and input—simultaneously improves the model's accuracy and guides the system towards preferred, livable, states that are characteristic of the kind of thing it is.

The key innovation of *active* inference lies in treating action selection as an inference problem, where policies (sequences of actions) are selected to minimize expected uncertainty (i.e., surprises in the future consequent on a policy). In order to handle the separation of temporal scales in real-world environments — and to balance present-moment expectations with future needs — the generative model is almost universally hierarchical, with higher levels encoding more abstract and longer-term predictions. For example, soundwaves can be abstracted into phonemes, which can be abstracted into syllables, and then into words, sentences, biographies, and so on (Baltzell et al., 2019; Dehaene et al., 2015; Ding et al., 2015; Taylor et al., 2015; Friston et al., 2024; Friston et al., 2017; George and Hawkins, 2009). This formulation allows for deep narratives about our experiences and our bodies across time, as well as goal-directed behavior, to emerge from the fundamental drive to maintain existence.

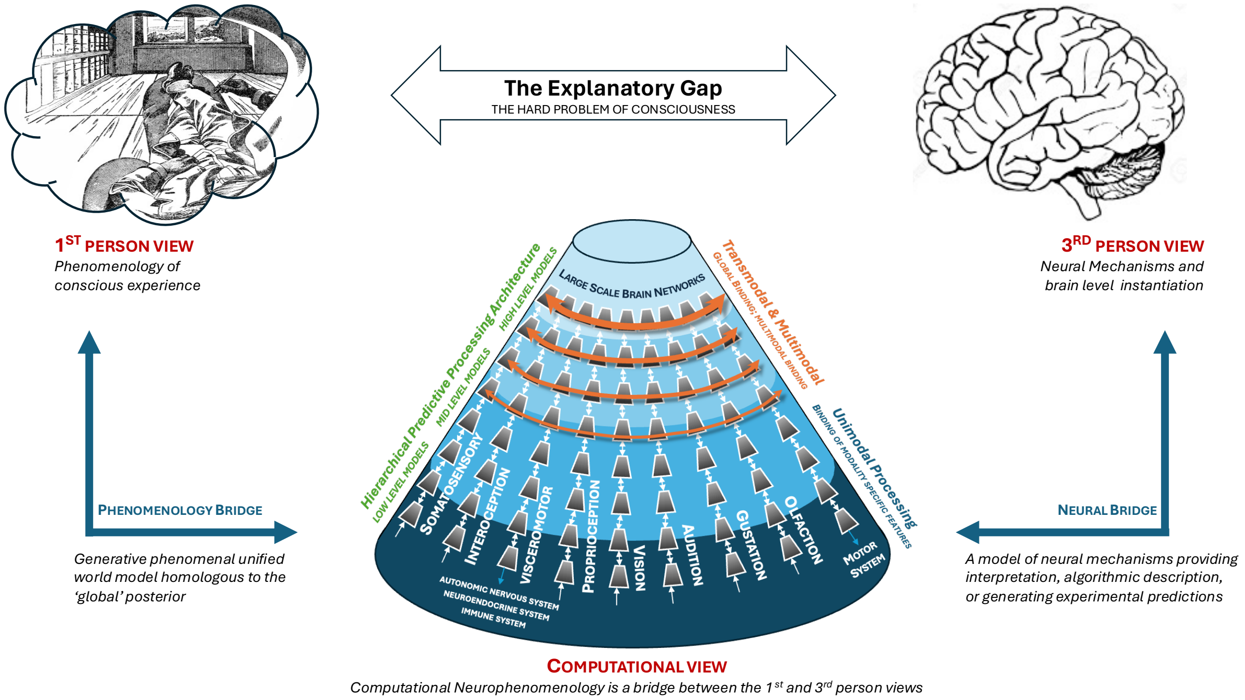
Finally, the prediction errors that report the degree of surprise — and thereby drive Bayesian belief updating at each hierarchical level — are precision-weighted (Feldman & Friston, 2010). Precision modulates the impact of prediction errors on belief updating and policy selection by controlling the gain on error units. High precision amplifies the influence of prediction errors, while low precision attenuates them. This precision-weighting mechanism allows the system to flexibly adapt to different contexts by modulating the balance between sensory evidence and prior beliefs. Put simply, we need to know what we know but also how confident we are about it. In some cases, we can trust our beliefs, in other cases we need to focus on learning something new from the world (Friston et al., 2015). On this reading, world models are 'precision engineered', where increasing the precision of certain prediction errors can be understood in terms of attending to their source.

One of the advantages of active inference is that it can act as a bridge between first and third-person approaches (cf. the *explanatory gap*, Levine, 1983). In other words, computational approaches like active inference offer a middle-way between subjective experience and neural mechanisms, providing mechanistic insight into both (see Figure 1). This approach is sometimes called *computational neurophenomenology* (Suzuki et al., 2022; Sandved-Smith et al., 2021; 2024; Ramstead et al., 2022) because it bridges subjective experience and objective neural processes within a single modeling framework. Specifically, it uses *generative models* to specify how the brain infers and constructs experiential content, allowing researchers to link changes in neural dynamics (the "algorithmic descriptions") to the qualities and structure of phenomenological experience. By systematically mapping both first-person reports and neural dynamics to underlying computational processes, we simultaneously gain an explanatory account of how subjective experience arises and a mechanistic understanding of how the brain implements it.

[1] Technically an upper bound on surprise or negative log evidence

Figure 1

Bridging the explanatory gap with computational neurophenomenology


The Explanatory Gap
THE HARD PROBLEM OF CONSCIOUSNESS
1ST PERSON VIEW
Phenomenology of conscious experience
3RD PERSON VIEW
Neural Mechanisms and brain level instantiation
PHENOMENOLOGY BRIDGE
Generative phenomenal unified world model homologous to the 'global' posterior
NEURAL BRIDGE
A model of neural mechanisms providing interpretation, algorithmic description, or generating experimental predictions
Hierarchical Predictive Processing Architecture
HIGH LEVEL MODELS
MID LEVEL MODELS
LOW LEVEL MODELS
LARGE SCALE BRAIN NETWORKS
Transmodal & Multimodal
GLOBAL BINDING; MULTIMODAL BINDING
Unimodal Processing
BINDING OF MODALITY SPECIFIC FEATURES
SOMATOSENSORY
INTEROCEPTION
VISCEROMOTOR
PROPRIOCEPTION
VISION
AUDITION
GUSTATION
OLFACTION
MOTOR SYSTEM
AUTONOMIC NERVOUS SYSTEM
NEUROENDOCRINE SYSTEM
IMMUNE SYSTEM
COMPUTATIONAL VIEW
Computational Neurophenomenology is a bridge between the 1st and 3rd person views

Note. This figure illustrates the explanatory gap between neural mechanisms and subjective experience. Hierarchical active inference (the cone in the middle) acts as a bridge between these two—first and third person—approaches to knowledge. The cone also provides a schematic overview of how a reality or world model can be constructed through a process of hierarchical precision-weighted prediction-error minimization (i.e., active inference). At the lowest level (dark blue), the organism encounters input from various systems, including the five senses as well as interoceptive, proprioceptive, visceromotor, immune, neuroendocrine, and gustatory systems. Through a continuous interaction — between top-down expectations and bottom-up prediction errors — the system constructs increasingly abstract and temporally deep representations giving rise to the self, world, thoughts, action plans, feelings, emotions, imagination, and everything else. As a primer for the next section, the cone also depicts how 'binding' may be occurring at various levels of the hierarchy, from low level features, to objects, to global multimodal and transmodal binding of the different parallel systems. Not depicted here is the fact that this hierarchical process is constantly tested and confirmed through action (e.g., top-down attention, physical movement, or reasoning).

Given the above, it appears that active inference can account for the capacity to simulate a pragmatic model of reality (i.e., a reality model that provides an epistemic field for adaptive action). Indeed, generating such a model is at the heart of active inference because without it the organism fails to anticipate preferred states and therefore maintain their boundaries — and bodies (Barrett, 2020). A growing evidence base suggests that active inference is, or at least could be, what the brain and body are doing (Walsh et al., 2020; Hohwy, 2013). Active inference therefore satisfies our first condition for consciousness, i.e., the generation of a world or reality model. This suggests that active inference may at least explain *what* is known or experienced. However, it does not yet explain the why or how of consciousness.

3. INFERENTIAL COMPETITION AND BAYESIAN BINDING

Any theory of consciousness must explain why we become aware of some phenomena but not others. Active inference and predictive coding have provided impressive accounts of how particular contents of consciousness are constructed (particularly in the visual stream, Peelen et al., 2024). But as yet, there is no generally accepted account of what determines the selective threshold required for awareness (Seth & Bayne, 2022; Baars, 2005; Kouider & Dehaene, 2007). Some informative efforts in this general direction exist (Saffron, 2020; 2022; Friston, 2018; Hohwy, 2012; Dołęga & Dewhurst, 2021).

Consider the familiar case of binocular rivalry (Breese, 1909; Tong et al., 2006). Here, a different image is presented to each eye at the same retinal location at the same time (e.g., a face is presented to the left eye, and a house is presented to the right eye). This results in a bizarre situation where the brain cannot seem to accept the fact of the two opposing visual realities. The resulting experience is a gradual switch between a house or a face, or a mix of the two. Such experiments highlight that some sort of selection process is taking place, which makes some configuration or interpretation of the senses conscious and not others (Hohwy et al., 2008; Hohwy, 2012). There are countless examples that raise a similar conundrum, including inattentional blindness (Mack, 2003; Kouider & Dehaene, 2007), visual and sensory illusions (Eagleman, 2008; Laukkonen & Tangen, 2017), as well as introspective, cognitive, and behavioral confabulations (Nisbett & Schachter, 1966; Nisbett & Wilson, 1977; Maier, 1931; Wegner, 2002; Weiskrantz, 1986).

While there are clearly some things that become conscious and others that do not, this pertains less to the harder problems of consciousness than one might think (Chalmers, 1995). Consider that regardless of which percept becomes conscious during binocular rivalry, we are always nevertheless (meta) aware of what enters our visual field. In other words, the *presence* of consciousness has not changed, only the contents. We can also be aware of the fact that our experience is changing from the face to the house, and even perhaps aware of why it is changing.

What is somehow central is therefore the fact that there seems to be a "space" wherein there is something it is like to feel, perceive, and crucially, know the contents of mind (Metzinger, 2020). Cleeremans et al (2020) put this same point differently: "...someone is aware of some state of affairs not merely when she is sensitive to that state of affairs, but rather when she *knows* that she is sensitive to that state of affairs." [our emphasis]. When encountering a visual illusion, rivalry, or an ambiguous stimulus, we seem capable of knowing that we are having such and such experience. In the parlance of phenomenology, we experience 'seeing' and phenomenological *transparency* gives way to *opacity* (Limanowski and Friston, 2018; Metzinger, 2003). This experiential space seems to be unified, coherent, and bound together: a conscious *gestalt*, as others have noted (Baars et al., 2013; Tononi et al. 2005, 2008). Hence, consciousness has a certain conclusive nature to it, as if the brain and body have found a global and unified affordance within which it can have a self, move about, attend to things, feel emotions; and crucially, keep the body alive (Barrett, 2020; Seth, 2013).

Here, we propose that the threshold for consciousness—what enters this epistemic field or reality model—is determined through a process of competition among possible explanations of the causes of one's sensations. Moreover, we suggest that the so-called binding "problem" may in fact be part of the "solution" as to what breaches the threshold for consciousness. That is, coherence and boundedness are a central criterion in winning the inferential competition. Metaphorically, the competition for consciousness has goalposts in the shape of coherence and unification. If an inference is incoherent with other parallel and hierarchically adjacent inferences throughout the system, then it is less likely to be selected[2].

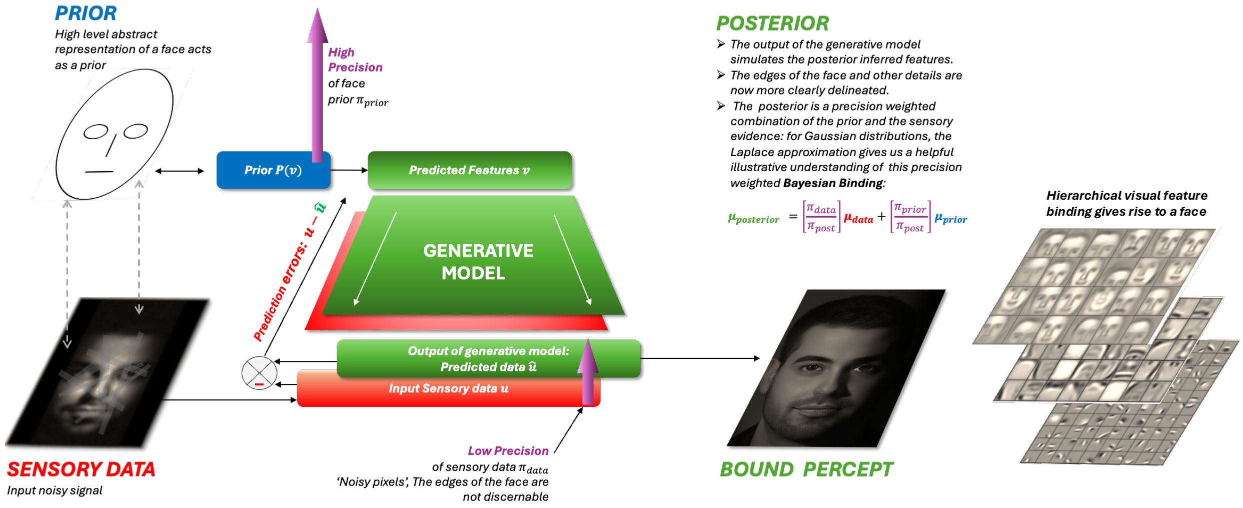
This *coherence criterion* also falls out naturally from a system that aims to reduce uncertainty or prediction-errors. Dissonance between inferences is equivalent to confusion—a generative model that does not parsimoniously explain the data. Such confounding explanations result in irreducible error propagation. The pressure for coherence is foregrounded by the fact that the remit of the reality model is to reduce uncertainty for adaptive action (Nave et al., 2020). If the epistemic field is internally incoherent, uncertainty accumulates as we evaluate policies or paths into the future, rendering action selection imprecise and their outcomes uncertain (i.e., surprising on average).

Specifically, we hypothesize that coherence and binding naturally fall out of a system that engages in hierarchical *Bayesian* inference (Knill & Pouget, 2004). What drives selection as to what gets bound into the field of experience is a precision-weighted competition between possible explanations for the causes of sensory data (i.e., a kind of competition for 'fame in the brain', Dennett, 1995; 2001). Crucially, what wins the precision-weighting competition is partially driven by the contents that best cohere with the existing reality model (i.e., priors), which provides the necessary constraints, or inducted biases (technically, empirical priors), for what can be assimilated into the epistemic field. In Bayesian terms, incoherent or incongruous data is either imprecise—in which case the associated prediction errors would be endowed with less precision—or unexplainable—in which case, precision weighting would preferentially select those data that can be explained. We illustrate this process of *Bayesian binding* using an example of micro-binding in a face percept (Figure 2). We argue that the same idea can be extended to macro-binding under the reality model.

[2] See for example the invisible gorilla effect (Simons & Chabris, 1999)

Figure 2

An example of “micro” binding for generating a face percept


PRIOR
High level abstract representation of a face acts as a prior
High Precision of face prior π_{prior}
Prior $P(v)$
Predicted Features v
GENERATIVE MODEL
Prediction errors: $u - \hat{u}$
Output of generative model: Predicted data $\hat{u}$
Input Sensory data u
Low Precision of sensory data π_{data}
‘Noisy pixels’, The edges of the face are not discernable
SENSORY DATA
Input noisy signal
POSTERIOR
BOUND PERCEPT
Hierarchical visual feature binding gives rise to a face

- The output of the generative model simulates the posterior inferred features.
- The edges of the face and other details are now more clearly delineated.
- The posterior is a precision weighted combination of the prior and the sensory evidence: for Gaussian distributions, the Laplace approximation gives us a helpful illustrative understanding of this precision weighted **Bayesian Binding**:

$$\mu_{posterior} = \left[\frac{\pi_{data}}{\pi_{post}}\right]\mu_{data} + \left[\frac{\pi_{prior}}{\pi_{post}}\right]\mu_{prior}$$

Note. This figure illustrates a simplified process of Bayesian binding in the context of face perception. The diagram shows how noisy sensory input is combined with prior expectations to produce a clear posterior representation under a generative model. Left: The sensory data shows a low-precision (noisy) input image of a face where details are not easily discernible. Top left: The prior is represented as a high-level abstract face shape, indicating the brain's pre-existing expectation of what a face looks like (inspired by Lee & Mumford, 2003). NB: In reality, the generative model has many levels, representing a continuous range of abstraction. Center: The generative model uses the prior P(v) to generate predicted features (v) that are combined with the sensory data (u) to produce prediction errors (u-$\hat{u}$), that together inform a posterior. Center Right: The posterior is the output of the generative model, showing a clearer, more detailed face image. This represents the brain's inference after combining prior expectations with sensory evidence. The equation illustrates a precision-weighted *Bayesian binding* process in a simplified unidimensional case assuming only Gaussian probability distributions. It shows how the posterior mean (μ_posterior) is a weighted combination of the prior mean (μ_prior) and the sensory data (μ_data), with weights determined by their respective relative precisions (π). This figure illustrates a key principle of Bayesian binding: a conscious percept or “thing” arises from the brain's attempt to create a coherent, unified explanation (the posterior) for its sensory inputs by combining them with prior expectations through hierarchical Bayesian inference. On the right, we also provide an intuitive monochrome visual illustration of feature binding in vision wherein low level visual feature patches are bound into face features like eyes, noses and mouths, and then how these features are bound into faces.

Bayesian binding also offers a novel description of *ignition* as defined in GNWT (Dehaene & Changeux, 2011; Dehaene et al., 2014; Friston et al., 2012). Ignition refers to the sudden and widespread activation of a coalition of neurons that “ignites” information into conscious awareness. In GNWT, this process is characterized by a nonlinear transition from local, specialized processing to global availability of information across the brain. According to Bayesian binding, the ignition threshold is driven by precision competition throughout the hierarchy, wherein precisions are also constrained by coherence (top-down) with the reality model (i.e., predicted precision[3] is higher if it has local and global coherence). Ignition,

[3] The factors that drive precision are manifold, including salience, top-down attention, context contingencies, coherence, uncertainty, confidence, reward, neuromodulators, and previous learning more generally. Moreover, top-down attention can “magnify” particular layers or contents within the hierarchy by increasing their relative (precision) weighting in the reality model (Feldman & Friston, 2010), e.g., by paying attention to the textured bark of a particular tree — in the context of a forest scene — the conscious gestalt will be modified by enhancing the details of the gnarliness of the bark. The low level sensory percepts occupy more ‘bandwidth’ in the epistemic field.

binding, and competition are hence subsumed within active inference (Whyte & Smith, 2021). They are each the natural consequence of a system that reduces uncertainty with sufficient complexity and depth.

A complimentary view (cf. Whyte & Smith, 2021; Whyte et al., 2024) proposes that consciousness arises specifically at the interface between continuous sensory perception and discrete, counterfactual policy selection processes. Here, conscious contents correspond to precise posterior beliefs about the hidden states of the world, body, or brain, which are temporally abstracted from immediate sensory fluctuations and sufficiently precise to drive action selection, including subjective reports. Thus, conscious states differ from unconscious ones in their capacity to inform discrete policy decisions, reflecting a computational balance between goal-directed (exploiting known information) and exploratory (resolving ambiguity and novelty) imperatives. An integrative perspective may be that *Bayesian Binding is a key mechanistic threshold between continuous sensory perception and discrete, conscious, and precise posteriors for counterfactual policy selection.* That is, Bayesian binding emphasizes that discrete and precise posteriors (Whyte & Smith et al., 2021) also require local *and* global coherence, which naturally drives the bounded and holistic nature of conscious experience.

The key takeaway here is that Bayesian binding is (in theory) at the heart of all experience, from unimodal processes, multisensory binding, through to global integration. That is, generating experience demands nested levels of binding, i.e., combining priors and sensory evidence into an approximate posterior through a hierarchical generative model, where that perceptual synthesis or combination rest sensitively on the precision or confidence afforded each level of processing (Friston, 2008; Hohwy, 2012; Hohwy, 2013). The insight is that the same basic mechanism operates at both micro and macro levels. For example, just as we infer a teapot to be a single 'thing' (made of handles, a hollow body, and a spout), we also take our whole field of experience to be a single thing, which binds together the ground, the sky, our bodies, other people, and everything else. We hypothesize that this global unified model, however minimal, is necessary but not sufficient for conscious experience. It is only when this global posterior is reflected back through the underlying hierarchy that the conditions for consciousness are met. As we will see, explaining the contents of consciousness is insufficient to capture the imminent and sense of *knowing* the contents, or *awareness* of them.

4. EPISTEMIC DEPTH AND AWARENESS

> *"...my own working hypothesis is that consciousness is our inner model of an "epistemic space," a space in which possible and actual states of knowledge can be represented. I think that conscious beings are precisely those who have a model of their own space of knowledge—they are systems that (in an entirely nonlinguistic and nonconceptual way) know that they currently have the capacity to know something."*[4]
>
> - *Metzinger, 2020*

[4] We generally agree with Metzinger's (2020; 2024) emphasis on an epistemic "space", though we prefer the term *field* because it highlights that the field and its contents are 'one' and always changing.

The word epistemic means ‘relating to knowledge’ and the word ‘depth’ refers to both intensity and the capacity to go below or beyond the surface (Oxford English, 1989). What we mean by the term epistemic depth is therefore a capacity or continuum (i.e., deepening) of knowing, or awareness, that can be more or less active (i.e., intense or clear). A state of low epistemic depth is one that involves unclear knowing, such as states of sleeping, dreaming, or mind-wandering, and a state with high epistemic depth is one that involves clear or intense knowing, such as mindful or highly aware states (Schooler, 2002; Schooler et al., 2011). As we will see, the intensity or clarity of knowing can also be directed at the knowing capacity itself, i.e., reflectively knowing that we know (Dunne et al., 2019; Josipovic, 2019). Our goal in this section is to deliver a conceptual sense of what we mean by epistemic depth. We then focus on providing a formalization of this idea in Section 5.

To add some phenomenological nuance to our construct we borrow the term *luminosity*, which appears often in the ancient discourses of contemplative traditions (Anālayo 2017), particularly in Mahayana and Vajrayana Buddhism (Williams, 2013) and Indian philosophy (Skorupski, 2012; Berger, 2015). For our purposes we define luminosity as the clarity or intensity of knowing or awareness within conscious experience. Connecting epistemic depth with luminosity has the particular benefit of avoiding an infinite regress: “as a source of light is never illuminated by another one…” (Bhartṛhari, 1963). Just as a light shining out of a lamp illuminates the objects but also the lamp itself, ideas of luminosity and self-reflective awareness often go hand in hand (Williams, 2013). Luminosity and recursion indicate that it is just ‘one’ knowing with varying degrees, just as a light varies in brightness, but is one light. For our purposes, luminosity provides a useful metaphor—with phenomenological resonance—for the graded nature or clarity of awareness that seems to be possible for conscious systems.

Now, returning to our construct of the reality model. In this context, luminosity is the degree to which the reality model (non-locally) knows itself. Within a hierarchical active inference system, the requisite sharing means that the reality model entails the inference, belief, or expectation, that it exists. By metaphor, it is as if the system’s output becomes another sensory modality that is recursively distributed back through all layers of the system. To provide a metaphor: When we speak out loud, we produce sounds and at the same time hear those sounds and their meaning (i.e., we hear our own voice and what we are saying). Therefore, our output (voice) is also our input (sound). We sequentially produce form (output) and then monitor the global context of that form (input) to ensure that our speech communicates a coherent stream of meaning[5]. There is a continuous ‘looping’ between what we create through action and what we perceive through the senses. Similarly, the key output of the inferential process in the brain is the construction of a reality model that allows us to survive (analogous to the voice). But this global reality model is also an input to the system and becomes part of the inferential process itself (analogous to the sound, cf. Figure 3).

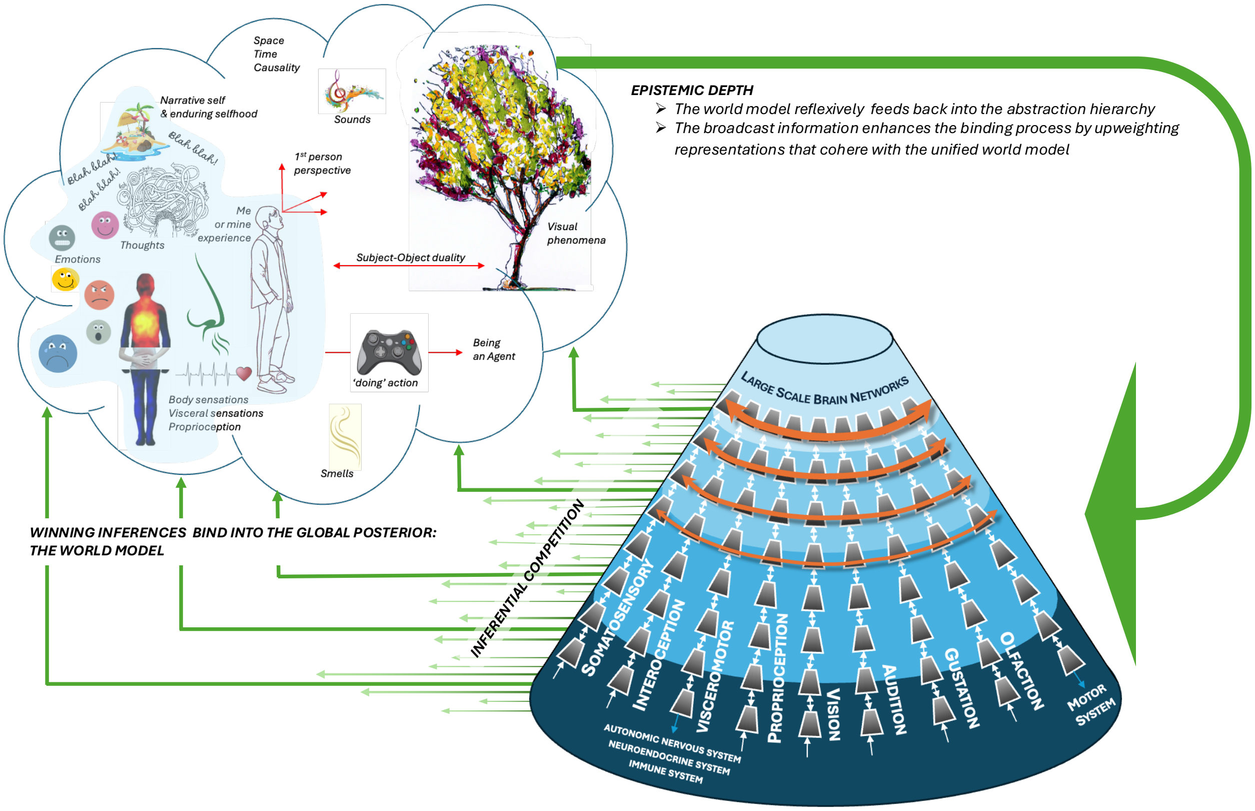
Consider that while particular contents of the reality model can be confirmed or disconfirmed by new evidence (e.g., transitions in binocular rivalry), the *existence of* the reality model is nevertheless receiving

[5] A more nuanced treatment of this analogy would call upon sensory attenuation; namely, the attenuation of the precision of the sensory consequences of self generated outputs. Following action, sensory attenuation is suspended so that we can attend to the consequences of what we have just done. This is clearly evinced in saccadic eye movements, where sensory attenuation is known as saccadic suppression, while the sensory attention — during successive fixations — underwrites epistemic foraging of a visual scene required to construct a perceptual gestalt.

continuous validation, regardless of the contents (e.g., all changes confirm that a reality model exists). Hence, the epistemic field is constantly evidencing its own existence (i.e., *field-evidencing*). Any action the organism takes—as little as a saccade, a thought, or a breath—confirms to itself that it (the model) exists. Indeed, all model (i.e., Bayesian belief) updates confirm it. Hence, the fact that the reality model exists becomes a precise inference that rarely loses the inferential competition.

Figure 3

Generating an epistemic field and its reflective sharing



Note. This figure illustrates the integration of information (operationalized by the hierarchical generative model (HGM)) into a reality model via nested Bayesian binding. The cone at the center illustrates a multi-tiered HGM structure with increasing levels of abstraction, from basic unimodal processes to abstract reasoning exemplified by large scale networks in the brain (Taylor et al., 2015). The cone includes feedforward and feedback loops throughout all layers. Increasing abstraction reflects increasing compression, information integration, temporal depth, and conceptualization (cf. Figure 1). A weighted combination of features across the hierarchy are combined or bound together via inferential competition (faded green arrows) to form a global posterior which is homologous to the reality model (the "conscious cloud" on the top left). This conscious cloud contains diverse perceptual, sensory, and conceptual elements, connected to corresponding hierarchical levels. Crucially, the reality model is recursively broadcast back throughout the hierarchy in the form of top-down predictions of both content and context (thick green arrow), where context is instantiated by predictions of precision. Crucially, predictions of precision weight the prediction errors that underwrite those predictions in a recursive fashion. This sharing of the reality model fine-tunes inference for binding by upweighting representations that cohere with it. We hypothesize that this recursion is the causal mechanism permitting epistemic depth (*the sensation of knowing*) because the information contained in the reality model loops back into the "conscious cloud" via the implicit abstraction hierarchy. Hence, the reality model contains information about the existence of itself. While the 'loop' is shown to and from the conscious cloud to illustrate the schema, computationally, all the recursion is within the feedback loops of the central cone structure: there is no dualism implied in this account.

5. HYPER-MODELS AND TINY CREATURES

Modeling *epistemic depth* in a rigorous way calls for a system that not only forms predictions about external states but also—crucially—*models its own modeling* recursively at a global scale. One way to pursue this is through what we term *hyper-generative models,* or *hyper-models* for short (Friston, 2010; Parr & Friston, 2018; Ramstead et al., 2022). In hierarchical active inference, each layer infers hidden causes in ascending degrees of abstraction. However, to capture epistemic depth, the architecture requires a *truly* higher-order (i.e., *hyper*) model that tracks how each layer's inferences and precision-weightings are being deployed system-wide. Formally, we can posit a hyper-parameter set Φ that encodes beliefs about *which* layers to trust more (or less) under different contexts, *how* strongly to up- or down-weight prediction errors, and *how* to orchestrate feedback loops across the entire network (Friston et al., 2017). Such a deeply global parameter permits a system to recursively "rework" and "rediscover" their own modelling processes, and thus become a truly agentic self-constructing and deconstructing system, reminiscent of the way humans can intentionally and radically change themselves given the right motivation and context.

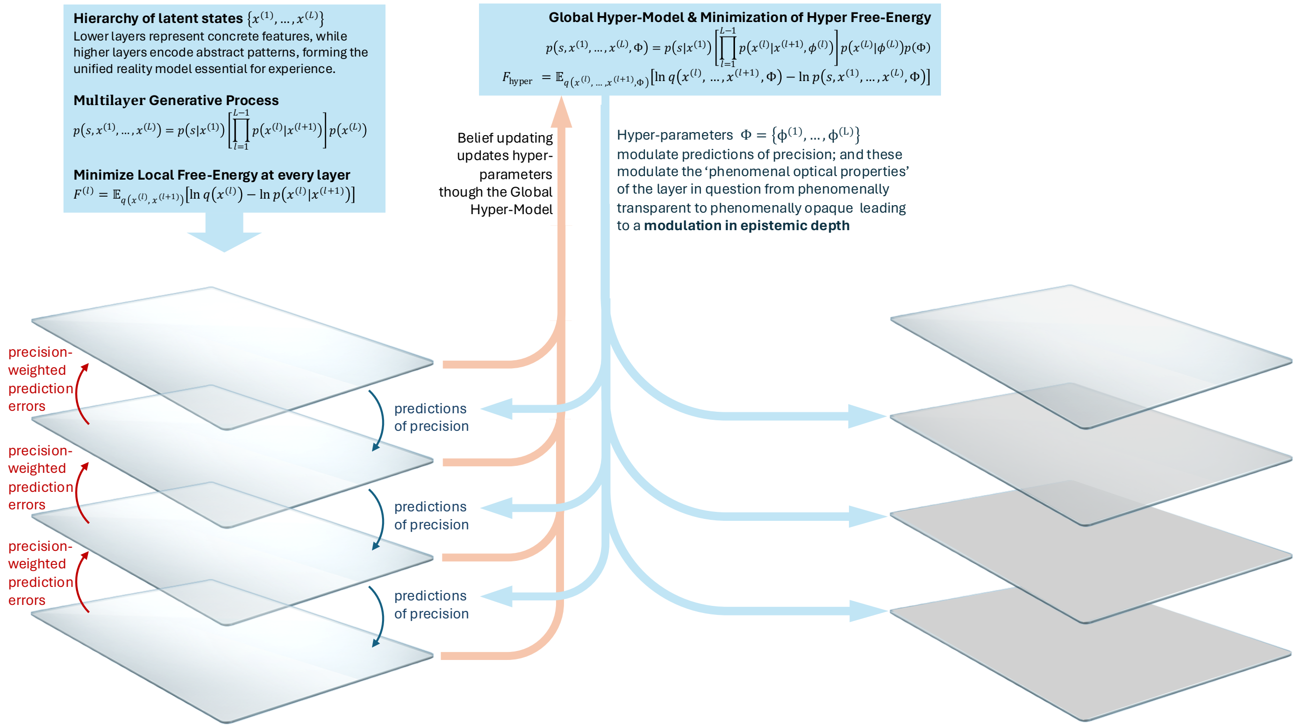
Crucially, hyper-parameters that contextualize belief updating—through descending predictions of precision to lower layers—update the (ascending) precision-weighted prediction errors that update the hyper-parameters that update (descending) predictions of precision. And so on, *ad infinitum*. It is this recursive aspect that equips belief updating with epistemic depth. In terms of phenomenal transparency and opacity, we can imagine each hierarchical layer as a type of glass that can change its optical properties (cf. Figure 4). In the setting of epistemic depth, descending predictions of precision render transparent panes of glass opaque, equipping the hierarchy with the ability to contextualize and select what is broadcast from one level to the next. In terms of a lamp illuminating itself, epistemic depth offers a very different picture: a picture more akin to a series of holographic screens (Fields et al., 2021; Fields et al., 2024) illuminating each other in their reflected light. This picture foregrounds the recursive, non-local and (self) reflective nature of epistemic depth.

Clarifying how a hyper-parameter set, Φ orchestrates the entire system is a challenge. One possibility is to define Φ within a factor-graph architecture that includes "hyper-nodes" encoding conditional beliefs about each sub-model's precision or reliability (Parr & Friston, 2018). These hyper-nodes would propagate top-down signals—precision updates, gating directives, or structural reconfigurations—to lower-level nodes, ensuring that each layer's inference is shaped by global meta-beliefs. Such a mechanism would allow simulation of when and how reflective broadcasts occur, enabling comparisons to neurophysiological data and refining our broader understanding of epistemic depth in biological and artificial systems.

Practically, this kind of architecture has proved useful in modelling brain responses (Iglesias et al., 2013), using a variant of predictive coding called the hierarchical Gaussian filter (Mathys et al., 2011). In computational neuroscience, minimal forms of epistemic depth have been used to illustrate attentional selection and the segregation of figure from ground (Kanai et al., 2015). Technically, the nonlocal aspect of epistemic depth inherits from the fact that the hyper-parameter—prescribing precisions at every level of the hierarchy—renders each level part of the hyper-parameter's Markov blanket (because they are all children of the hyper-parameter). This mandates recursive message passing between the Bayesian beliefs over hyper-parameters and all levels, in which descending predictions of precision are reflected back in the form of a prediction error over precision. See Kanai et al., (2015) for the functional form of these second-order prediction errors in the context of predictive coding architectures.

Figure 4

Epistemic depth as hyper-modeling

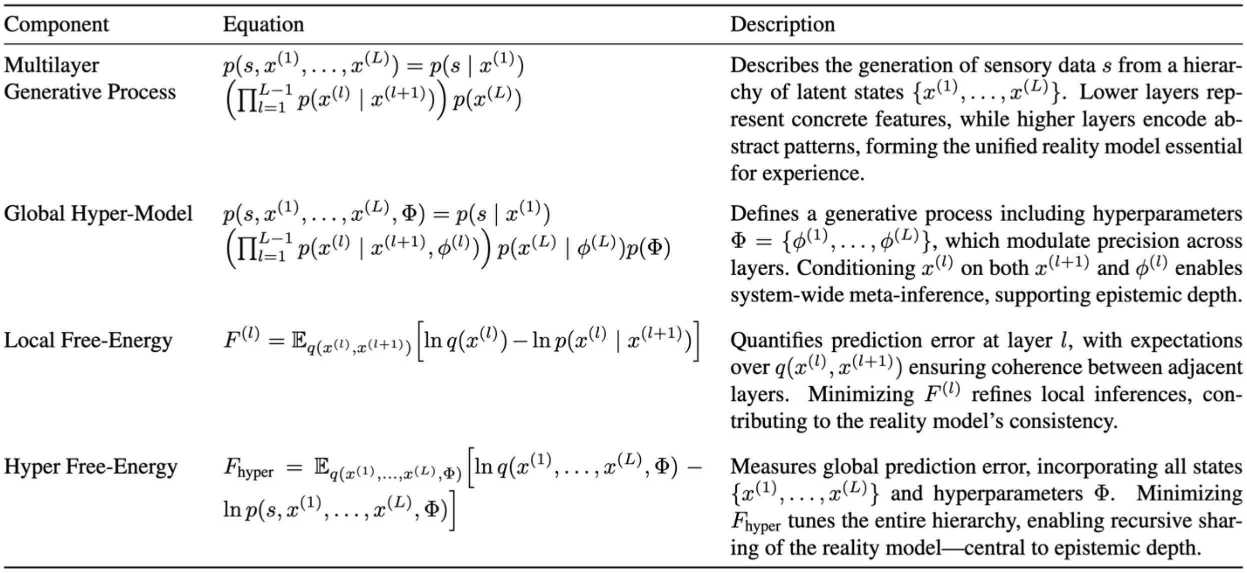

Hierarchy of latent states $\{x^{(1)}, \dots, x^{(L)}\}$
Lower layers represent concrete features, while higher layers encode abstract patterns, forming the unified reality model essential for experience.
Multilayer Generative Process
$p\left(s, x^{(1)}, \dots, x^{(L)}\right) = p\left(s|x^{(1)}\right)\left[\prod_{l=1}^{L-1} p\left(x^{(l)}|x^{(l+1)}\right)\right] p\left(x^{(L)}\right)$
Minimize Local Free-Energy at every layer
$F^{(l)} = \mathbb{E}_{q\left(x^{(l)}, x^{(l+1)}\right)}\left[\ln q\left(x^{(l)}\right) - \ln p\left(x^{(l)}|x^{(l+1)}\right)\right]$
Global Hyper-Model & Minimization of Hyper Free-Energy
$p\left(s, x^{(1)}, \dots, x^{(L)}, \Phi\right) = p\left(s|x^{(1)}\right)\left[\prod_{l=1}^{L-1} p\left(x^{(l)}|x^{(l+1)}, \phi^{(l)}\right)\right] p\left(x^{(L)}|\phi^{(L)}\right) p(\Phi)$
$F_{\text{hyper}} = \mathbb{E}_{q\left(x^{(l)}, \dots, x^{(l+1)}, \Phi\right)}\left[\ln q\left(x^{(l)}, \dots, x^{(l+1)}, \Phi\right) - \ln p\left(s, x^{(1)}, \dots, x^{(L)}, \Phi\right)\right]$
Belief updating updates hyper-parameters though the Global Hyper-Model
Hyper-parameters $\Phi = \{\phi^{(1)}, \dots, \phi^{(L)}\}$ modulate predictions of precision; and these modulate the 'phenomenal optical properties' of the layer in question from phenomenally transparent to phenomenally opaque leading to a **modulation in epistemic depth**
precision-weighted prediction errors
predictions of precision
precision-weighted prediction errors
predictions of precision
precision-weighted prediction errors
predictions of precision

Note. This diagram illustrates the abstraction hierarchy of features as being composed of layers of 'smart' glass which have the property of being able to change from being transparent (e.g., the layers on the left) to more or less opaque (e.g., the layers on the right) analogizing *phenomenal* transparency or opacity of that layer. By analogy, when a pane of glass is opaque, the contents of our consciousness are known (such as being aware of the feeling of wearing a shirt). On the other hand, when it is transparent, we do not notice the shirt—like looking through a clean window where we do not notice the glass. To account for this within hierarchical active inference, we propose the following: The (local) free energy of every layer of the multilayer generative process is minimized in the usual way, but as a crucial extension, global free energy is minimized in the context of a Global Hyper-Model which includes a set of hyperparameters $\Phi = \{\phi^{(1)}, \dots, \phi^{(L)}\}$ that control predictions of precisions at every layer. These hyperparameter controlled precision modulations can be said to regulate the 'phenomenal optical properties' of the layer in question from phenomenally transparent to phenomenally opaque leading to a fully endogenously determined modulation of epistemic depth globally. We unpack this further below and provide details in Table 1.

Unlike parametric depth, a hyper-model is modeling the very shape of its hierarchy and updating it in real-time. Parametric depth is often implemented as localized loops—between one layer above another (Sandved-Smith, 2021; 2024), implementing a second-order inference about attention, or about preference precision. One can extend this idea to multiple layers, but typically it is demonstrated with a single or small number of layers. Epistemic depth goes beyond local second-order inferences, implying a *globally consistent* sense that "I (the system) have a multi-tier generative model, and I know how to deploy the right precision in each tier—thus I *know* what I know." This kind of epistemic depth is system-wide: it is not just "what am I attending to?" but "how do all these layers of inference contextualize each other in a deep (hierarchical) sense?”. Whereas parametric depth can be instantiated with a few carefully chosen parameters: e.g., “likelihood precision” or “policy precision” (Allen et al., 2019; Hesp et al., 2019; Parr and Friston, 2017, 2019; Schwartenbeck et al., 2015; Smith et al., 2019), epistemic depth is about the *entire deep generative model* being “aware” of how it’s orchestrating priors, transitions, preferences, timescales, and so on. Nevertheless, parametric and epistemic depth are clearly compatible—epistemic depth sketches the ‘big-picture’ of global awareness, while parametric depth is a mechanism for implementing meta-inference in a hierarchical generative model that has close connections to metacognition and the higher-order thought theory (Fleming, 2020; Fleming et al., 2012).

Table 1

Towards a Formal Model of Epistemic Depth



Component	Equation	Description
Multilayer Generative Process	$p(s, x^{(1)}, \ldots, x^{(L)}) = p(s \mid x^{(1)}) \left(\prod_{l=1}^{L-1} p(x^{(l)} \mid x^{(l+1)})\right) p(x^{(L)})$	Describes the generation of sensory data s from a hierarchy of latent states $\{x^{(1)}, \ldots, x^{(L)}\}$. Lower layers represent concrete features, while higher layers encode abstract patterns, forming the unified reality model essential for experience.
Global Hyper-Model	$p(s, x^{(1)}, \ldots, x^{(L)}, \Phi) = p(s \mid x^{(1)}) \left(\prod_{l=1}^{L-1} p(x^{(l)} \mid x^{(l+1)}, \phi^{(l)})\right) p(x^{(L)} \mid \phi^{(L)}) p(\Phi)$	Defines a generative process including hyperparameters $\Phi = \{\phi^{(1)}, \ldots, \phi^{(L)}\}$, which modulate precision across layers. Conditioning $x^{(l)}$ on both $x^{(l+1)}$ and $\phi^{(l)}$ enables system-wide meta-inference, supporting epistemic depth.
Local Free-Energy	$F^{(l)} = \mathbb{E}_{q(x^{(l)}, x^{(l+1)})} \left[\ln q(x^{(l)}) - \ln p(x^{(l)} \mid x^{(l+1)})\right]$	Quantifies prediction error at layer l, with expectations over $q(x^{(l)}, x^{(l+1)})$ ensuring coherence between adjacent layers. Minimizing $F^{(l)}$ refines local inferences, contributing to the reality model’s consistency.
Hyper Free-Energy	$F_{\text{hyper}} = \mathbb{E}_{q(x^{(1)}, \ldots, x^{(L)}, \Phi)} \left[\ln q(x^{(1)}, \ldots, x^{(L)}, \Phi) - \ln p(s, x^{(1)}, \ldots, x^{(L)}, \Phi)\right]$	Measures global prediction error, incorporating all states $\{x^{(1)}, \ldots, x^{(L)}\}$ and hyperparameters Φ. Minimizing F_{hyper} tunes the entire hierarchy, enabling recursive sharing of the reality model—central to epistemic depth.

Note. A simplified description of each component is as follows. Multilayer Generative Process: This is the standard hierarchical formulation found in active inference and predictive coding. Global Hyper-Model: The novelty here is to include an extra “layer” of inference that regulates how each level in the hierarchy should be trusted—by updating the hyper-parameters Φ that set precision and weighting rules across levels. This reflective control is our formal analog of epistemic depth. Local and Hyper Free-Energy: The free-energy formalism provides a way to quantify “prediction error” at both local and global levels, where updates are shared recursively between local and global levels. The local free-energy drives short-term, layer-by-layer inference, whereas the hyper free-energy ensures that the whole hierarchy (including its meta-parameters) is optimized. These local and global processes may be what give rise to the sense that there is a difference between awareness and its contents: The global level (awareness) always seems to track the functioning and structure of other, localized loops (contents) in the context of the whole.

It also remains an open question where, along the phylogenetic continuum, genuine *epistemic depth* begins to appear. Biologically, all living systems engage in some form of homeostatic regulation, and many (e.g., bacteria) exhibit simple feedback loops. However, the *simplest formal demonstration of epistemic depth is probably a two(+)-layer active-inference system that:*

1. Infers external states (a minimal world model - the “what” of consciousness)
2. Maintains a meta-layer that infers “confidence about those inferences” (minimal competition between possible interpretations of the causes of sensation)

3. Reflectively modifies the lower-level inference from the meta-layer's vantage, creating a closed loop of self-modeling (minimal epistemic depth)[6]

Even such a toy system can, in principle, encode a rudimentary "knowing that it knows." This simple setup forms a minimal demonstration of *epistemic depth*: the global loop includes not just a model of the world, but also a real-time model of *how* it is modeling the world (cf. Parr & Friston, 2018; Sandved-Smith et al., 2021). However, unlike reflecting on how one's mind works (in the way we are doing here), rudimentary forms of epistemic depth are exceptionally basic: they involve minimal meta-inference about a system's own predictive processes, absent any richer conceptual or introspective dimension. In this sense, our model seems to suggest that consciousness clearly precedes introspection or complex metacognition, at least the kind that we usually associate with those terms. Even a tiny agent can incorporate ongoing feedback from within its own inferential machinery, linking the "sense of the world" with a subtle, self-revising "sense of itself *as* the world." Self-modeling proper (i.e., knowing what kind of thing one is) would be, under this view, a much later development.

In nature, many single-cell organisms already show *proto*-forms of "self-measurement" of intracellular states, but whether that amounts to a reflective *awareness* is debatable (Fields & Levin, 2022; 2023). One could argue that very small multicellular creatures or tiny insects—e.g. parasitic wasps (*Megaphragma mymaripenne*), fruit-fly larvae (*Drosophila melanogaster*), or nematodes like *C. elegans*—start to approach the complexity needed to implement the minimal hyper-model. These small but highly integrated nervous systems can tune sensory signals, modulate action policies, and reconfigure local loops via neuromodulators, suggesting partial analogs of top-down reweighting (Marder, 2012). Whether these are enough for a *globally coherent* "knowing that they know" remains unknown, but they are prime targets for studying borderline cases of reflective, multi-level organization in living systems[7]. As a practical matter, it may be that only once a creature devotes sufficient neuronal (or computational) resources to hierarchical modelling and meta-inference do we see a clear approximation of *epistemic depth* in the sense required here (Friston, 2018). Nonetheless, tracking how progressively complex nervous systems—beginning even with tiny arthropods—handle global precision control may shed light on how minimal systems might, at least in principle, exhibit the core properties of epistemic depth.

6. ATTENTION AND METACOGNITION

At this stage it is useful to make explicit several levels of awareness according to our view: First, there are those inferences which have low precision and lose the competition for binding into the reality model. These inferences remain transparent and cannot be introspected at that moment. Second, there are inferences that have high enough precision and sufficient coherence with the epistemic field to win the inferential competition for binding (cf. Figure 2). These contents are at least subtly or barely known, but

[6] NB: *reflexively modifying the lower-level inference from the meta-layer's vantage*—describes a *closed, self-revising loop* that can in principle involve *all* the agent's inferences, not just a single parameter. Once the meta-layer's beliefs about "how I'm doing" feed back into *the entire inference process*, you get a broader or "global" loop that supports a rudimentary "I know that I know."

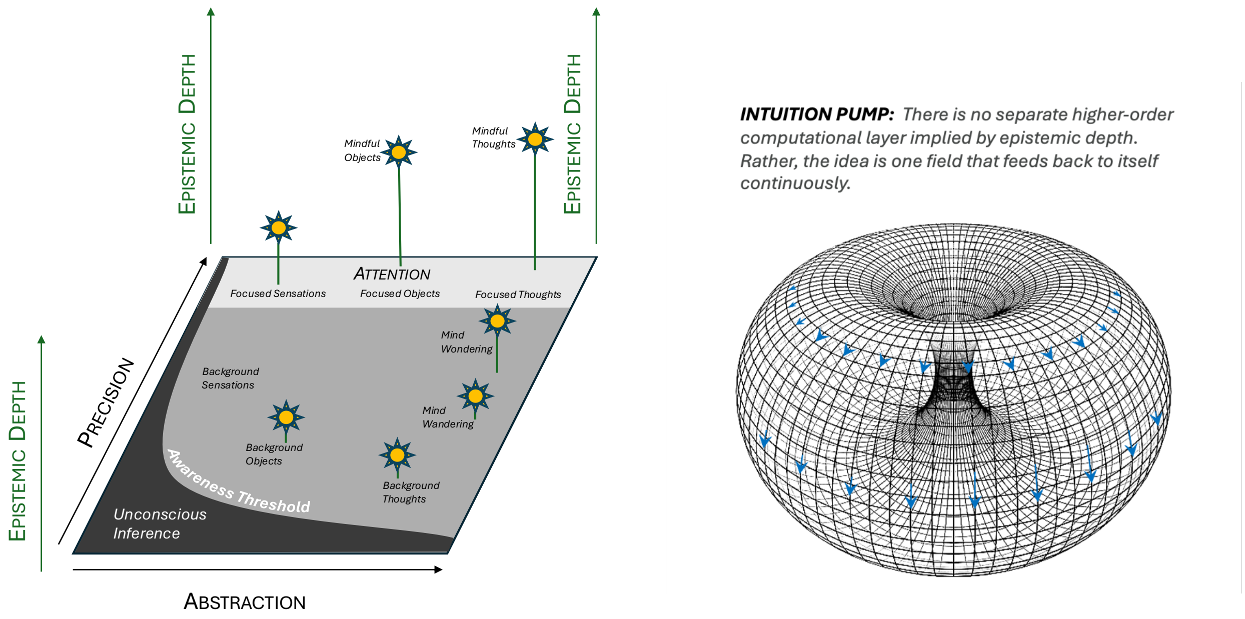
[7] We remain cautiously open minded about epistemic depth in the plant world.

we may not be explicitly 'knowing that we know' (e.g., the periphery of our attention, or the shirt that we are wearing). Whether content is 'luminously' aware depends upon epistemic depth, i.e., *hyper-modeling* (cf. Figure 3). For example, subliminal priming (Ansorge et al., 2014; Elgendi et al., 2018) occurs when input has sufficient precision to entrain hierarchical processing, but not enough to win the inferential competition for binding into the reality model. Truly subliminal information (e.g., the stages of processing underlying binocular rivalry or a visual illusion) cannot be introspectively accessed. Our framework also implies that due to Bayesian binding, if input is incoherent with one's current reality model, then even a precise bottom-up signal might not be noticed (e.g., inattentional and change blindness, Simons, 2000; Simons & Levin, 1997). An example of subtle knowing occurs when we find that we have driven our car (or walked) to our destination without being explicitly aware of the journey because we are busy mind-wandering or listening to a podcast (i.e., *relative* blindsight, Lau & Passingham, 2006). Here, the sensory and motor data associated with walking and driving are clearly part of the reality model, but they lack epistemic depth. They suffice for adaptive action, but lack the ingredients for awareness.

A crucial point to appreciate is that any content in the reality model can become the object that is explicitly and luminously known (cf. Figure 5): Hyper-modeling can explicitly track the structure and precision of any layer of inference. This includes metacognition, attentional processes (i.e., model simplifications or precision updates), and even the self (Dahl et al., 2015). We can be aware that we are thinking about thinking, or mindful of attentional states (Lutz et al., 2015). Hence any content that is at one point transparent (assumed, unknown) can become opaque (unassumed, known; Metzinger, 2003).

Figure 5

Epistemic depth as (partially) orthogonal to the precision-weighted abstraction hierarchy



INTUITION PUMP: *There is no separate higher-order computational layer implied by epistemic depth. Rather, the idea is one field that feeds back to itself continuously.*

Note. This three-dimensional model illustrates the relationships between abstraction (horizontal axis), precision (diagonal axis), and epistemic depth (vertical axis). Various cognitive states are mapped onto this space, with sensations, objects, and thoughts varying in their place within the precision-weighted abstraction hierarchy. Star-like symbols represent different conscious states, with their height indicating the degree of epistemic depth. In the bottom-left corner (dark gray), a process of unconscious inferential competition unfolds until an awareness threshold is passed (i.e., binding into the reality model). Within the space of awareness, ‘attention’ states (light gray) are simplified or focused reality models at different levels of abstraction. Mindful states are positioned higher on the epistemic depth vertical axis, suggesting increasingly clear ‘knowing of what is known’. For example, thinking is shown at various levels of epistemic depth, illustrating how the same cognitive process can vary in luminosity (e.g., from mind wandering, to mind “wondering” [intentionally allowing the mind to travel, Schooler et al., 2024], to mindful thoughts). The figure also shows broadly how targets of attention (high precision), but also phenomena in the periphery (relatively low precision), can change depending on the degree of epistemic depth. The toroidal figure on the right aims to provide a feeling or intuition for the way that epistemic depth works—it is not a separate thing but a continuous global sharing of information by the system with itself.

Within HOT theories of consciousness, metacognition is thought to be a necessary capacity for conscious experience (cf. Cleeremans et al., 2020; Shea & Frith, 2019, for reviews). In recent formulations, the kind of metacognition discussed is somewhat different to the familiar notion of metacognition as ‘thinking about thinking’ (Fleming & Lau, 2014). It is a more subtle kind of metacognition: a subpersonal or implicit ‘sensitivity to sensitivity’ (Cleeremans et al., 2020; Lau, 2022). Epistemic depth also involves a kind of ‘sensitivity to sensitivity’ in order to ‘know what we know’. However, this is achieved through recursion not metacognition. There is an element of re-representation here, but it is not hierarchically situated, it is recursive, like a new sensory modality. It is more akin to a *revelation*, where one’s epistemic state is continually revealed to oneself. Under this view, even complex metacognition can potentially be unconscious or conscious depending on epistemic depth (Kentridge, 2000). As noted by Koriat and Levy-Sadot (2000, p. 198), “... if metacognitive monitoring is defined as knowledge about one’s own knowledge, there is no a priori reason for denying the possibility that such knowledge might also be implicit and unconscious.”

We can speculate a step further. So-called System 2 processes (Kahneman, 2011) characterized by analytic, effortful, and linear thinking and reasoning, also depend on epistemic depth for awareness. This is consistent with findings wherein problems that seem to necessitate analytic processing are often solved through unconscious processes, i.e., sudden insights (Metcalfe & Wiebe, 1987; Laukkonen et al., 2023; Patel et al., 2019; Webb et al., 2018). Indeed, some of the greatest breakthroughs in science and mathematics have happened on the back of Eureka moments, where deep analytic processes continue their work below awareness (Salvi et al., 2024; Ovington et al., 2018; Kounios & Beeman, 2018). Likewise, we can be absorbed in complex analytic work akin to a flow state (Dietrich, 2004; Marty-Dugas et al., 2021; Parvizi-Wayne et al., 2024), with only subtle awareness of what we are doing, as in programming or scientific writing. Or we can be explicitly and luminously aware that we are thinking about thinking, and share the curious phenomenology of the experience with others. In sum: just as we can be aware (or not) of the low-level soft and woolly textures of a cozy jumper, we can be aware (or not) of abstract thinking, metacognition, and reasoning. Awareness depends on epistemic depth, not configurations of contents or the degree of analytic processing.

7. SLEEP AND LUCIDITY

Detailed applications of our theory to sleep states will be the subject of future work, but here we review them briefly. In non-rapid eye movement (NREM) sleep, particularly deep slow-wave sleep, the reality model becomes drastically simplified—the precision of sensory input is low. This results in a minimal, or transiently absent, reality model. Moreover, reflective broadcasting of the reality model throughout the hierarchy is massively diminished, associated with a state of low epistemic depth (i.e., low precision and low recursion) and hence reduced awareness or unconsciousness. This is supported by a wealth of evidence indicating that temporally deep processing and long-range functional connectivity and feedback processes are reduced during deep sleep (Massimini et al., 2005; Horovitz et al., 2009; Nir et al., 2011; Esser et al., 2009; Kakigi et al., 2003; Tagliazucchi & van Someren, 2017; Tononi & Massimini, 2008; Mashour & Hudetz, 2018; Laureys, 2005; see also vegetative states, Boly et al., 2011).

On the other hand, during rapid-eye movement (REM) sleep and dreaming, the reality model is richer and more complex. Here, some hierarchically deep and recurrent processing continues to occur sufficient for binding of a reality model, creating unified (albeit unusual) percepts and narratives that may lack the sense of logic or reason associated with truly high-order (e.g., pre-frontal) processes (Maquet et al., 1996). However, epistemic depth remains relatively low, resulting in a lack of awareness (i.e., lucidity) that one is dreaming—*we do not know that we know*. This is evidenced by a relatively stronger breakdown in abstract, temporally deep processing during NREM sleep (Wilf et al., 2016; Strauss et al., 2015; Massimini et al., 2005; Hayat et al., 2022) and the maintenance of some widespread connectivity during REM sleep, which accords with the general notion that REM is a kind of hybrid of NREM and wakefulness (Braun et al., 1997; Hobson & Pace-Schott, 2002; Nir & Tononi, 2010; Hayat et al., 2022).

Lucid dreaming offers a particularly intriguing case. In a lucid dream, one becomes aware that they are dreaming, often gaining some degree of control over the dream narrative (Saunders et al., 2016). Within our framework, lucid dreaming is a state where epistemic depth increases significantly compared to typical (non-lucid) REM. The boost in epistemic depth allows the dreamer to recognize the current state as a dream—there is a partial reactivation of the mechanisms that support reflective awareness in waking consciousness. This is consistent with practices during the day that support the emergence of lucidity at night, such as reality monitoring (Loo & Cheng, 2022) and mindfulness (Stumbrys et al., 2015; Stumbrys & Erlacher, 2017), both of which increase epistemic depth. There is also preliminary evidence that lucid dreaming is associated with re-activation of prefrontal brain regions (Baird et al., 2018; 2019; Vos et al., 2009; Dresler et al., 2012), which possess the widespread connectivity that may be important for reflective broadcasting (Dehaene et al., 2014; Miller & Cohen, 2001; Baird et al., 2018).

Finally, there is the even rarer possibility of lucid dreamless sleep (Windt et al., 2016; Thompson, 2015)—an awareness without any dream content during deep stages of sleep. Sometimes termed ‘clear light sleep’ in translations from Tibetan Buddhist works (Alcaraz-Sanchez, 2023), lucid dreamless sleep can be a spontaneous occurrence but is also an intentional practice in some contemplative circles (Thompson, 2015; Windt, 2020; Holecek, 2020). Descriptions of such ‘pure’ awareness during sleep go back at least as far as the Upanishads (classical Indian spiritual texts) where it is known as *sushupti* (Sanskrit: सुषुप्ति, Alcaraz-Sanchez, 2023). Interestingly, this state is also commonly characterized by ‘clarity’ and ‘luminosity’ (Padmasambhava & Gyatrul, 2008). Within our framework, lucid dreamless sleep is a state of very low abstraction but high epistemic depth: There are no constructed contents of consciousness and only epistemic depth (i.e., luminosity) remains, resulting in an experience of an aware but empty epistemic field.

8. MEDITATION & MINIMAL PHENOMENAL EXPERIENCE

It can be informative to consider how active inference can now accommodate some altered states of mind and consciousness, particularly those that can occur during long term meditation (Lutz et al., 2019; Pagnoni et al., 2019; Laukkonen & Slagter, 2021; Deane et al., 2020; Prest & Berryman, 2024; Berkovich-Ohana et al., 2024). We consider such broad applications of any theory of consciousness crucial: If the theory only provides a narrow window on a subset of conscious experiences, it is clearly unable to mirror the complex, multidimensional, and flexible nature of consciousness. A theory that overlooks such states is therefore at risk of missing the forest for the trees. Yet we also acknowledge that meditation states and psychedelic states (discussed later) are nebulous, and measuring and mapping them rigorously is notoriously difficult. But there is cause for optimism: Recent decades have provided a growing evidence-base of neurophenomenological data permitting a rapidly growing triangulation of these unusual—yet ancient and widespread—conscious states (Lutz et al., 2015).

We take a particular interest here in modeling what is known as *minimal phenomenal experience* (MPE; cf. Windt, 2015; Metzinger, 2020; 2024; Gamma & Metzinger, 2021; Woods et al., 2023; 2024; Dor-Ziderman et al., 2013; Ciaunica & Crucianelli, 2019). The philosophical idea is that the best model of consciousness is the simplest one: an explanation that takes the most basic or 'minimal' form of awareness as its target. This approach aims to avoid conflating consciousness *as such* with particular expressions of it, including self-hood, agency, time, or a first-person perspective (Metzinger, 2020; 2024). The best examples include pure, or contentless awareness experiences, lucid dreamless sleep, or other minimal non-dual awareness events reported by many contemplative traditions (Thompson, 2015; Hanley et al., 2018; Josipovic, 2019; Laukkonen & Slagter, 2021).

Some important groundwork already exists (cf. Metzinger, 2020; 2024; Josipovic, 2019). According to Metzinger, MPE may correspond to the phenomenology of "tonic alertness" — a state of bare wakeful awareness without any specific content. It is an abstract, contentless experience of "openness" and epistemic potential—a non-egoic representation of the mere capacity for knowledge and perception (i.e., the epistemic space). It is also luminous, "*...clarity inseparable from emptiness*" (Lingpa, 2014, pp. 14–15, quoted in Metzinger, 2020). Our view is thoroughly in agreement with Metzinger's general phenomenological characterization. However, we also agree with Josipovic (2019) that the recursion of non-dual awareness is *sui generis*, a unique and holistic capacity to non-conceptually become aware of the reality model. But crucially, this recursion does not depend upon any particular contents, and is not abstract. The reality model, through recursive sharing of its own structure and weighting rules (i.e., hyper-modeling), can know-itself continuously and simultaneously as both the content and the content that is known.

Formally, we propose that MPE occurs when: *epistemic depth is maximally high,* and *the reality model is contentless* (minimal precision across the abstraction hierarchy). Due to recursive broadcasting, this results in the recursion (i.e., luminosity) becoming the dominant input of the reality model (i.e., perpetually winning the inferential competition). This results in a kind of reflective recursion—awareness of awareness. But it is misleading to imagine this as dualistic or something that takes time. The input to the reality model, the reflective sharing, and inferential competition, are all co-occurring in the system. They are all part of one continuous process. It is a bit like acoustic feedback: the sound from a loudspeaker reenters the microphone and forms a perpetual loop. What emerges is what we can poetically call a "beautiful" rather than “strange” loop (Hofstadter, 2007); a kind of toroidal (cf. Figure 4) epistemicity that arises out of the global function rather than specific informational content or meta-representation.

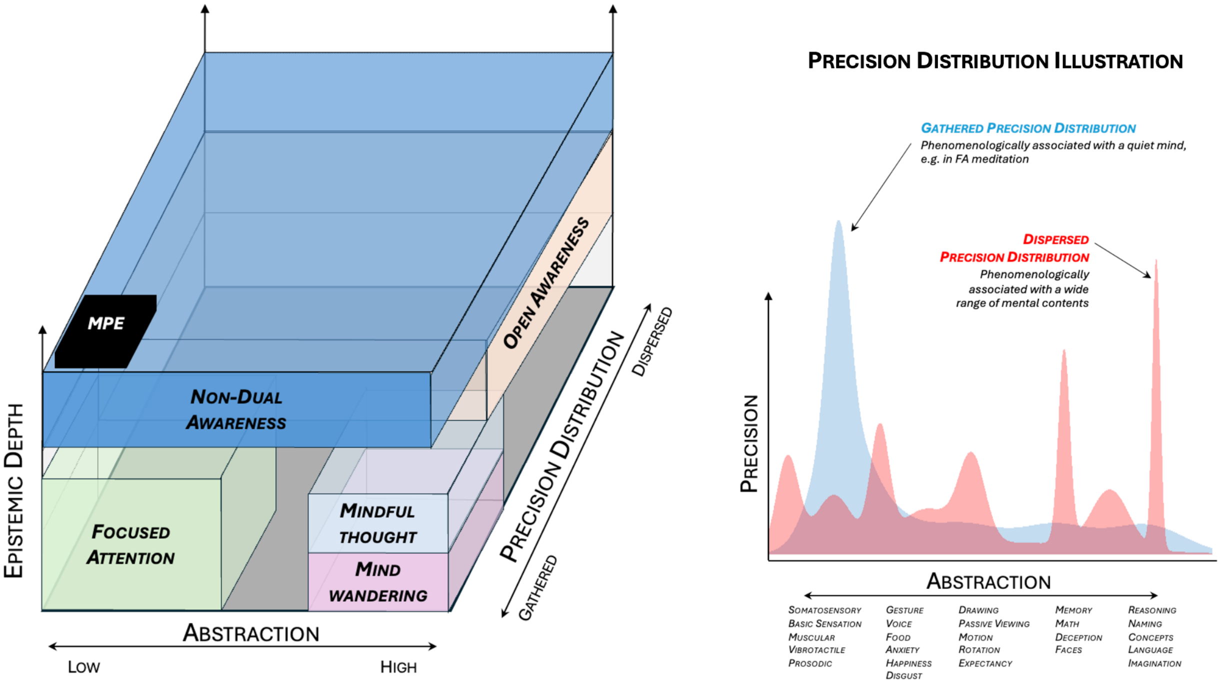
Metaphorically, it is as if the system is ‘focused’ on its own capacity to know. But as discussed earlier, *focus* is really *simplification*; and *knowing* is really *recursion*. Hence, if the reality model is simple enough that the only signal that wins the inferential competition is the recursion itself, then the recursion continuously shares itself with itself. This maps onto the phenomenal character of MPE (Metzinger, 2020), as luminous, simple, singular, non-dual, and true (i.e., precise). In the context of some minimal affect (i.e., *approximate MPE*), it also makes sense that it is blissful—there may be one affective signal that is almost perfectly explained by the model, hence uncertainty and associated tensions are exceedingly low compared to ordinary life. Moreover, by making bliss, joy, or happiness, the focus (i.e., high precision content in a simplified reality model), one can entrain a “beautiful loop” of sustained positive affect and awareness, analogous to the high pitched whistle of acoustic feedback described above, and reminiscent of what Buddhist’s call *Jhānas* (Hagerty et al., 2013; Laukkonen et al., 2023; Sparby & Sacchet, 2024).

It is unsurprising that the most truly minimal versions of MPE occur during deep sleep where other contents are the least active and recursion can dominate the reality model, and that such experiences might be more common among contemplatives who train in mindfulness and open awareness. Increasing epistemic depth during the day may reasonably be expected to entrain a habit of recursion such that it occurs during sleep. Similar to the way that we may dream about the events of our day; by habitually becoming aware of our reality model (conscious gestalt), we may increase inertia for reflective (non-dual) awareness during the night. In Figure 6, we extend our framework beyond MPE to other meditative states, including focused attention, open awareness, and non-dual awareness (Sparby et al., 2024; Dahl et al., 2015; Lutz et al., 2017; Slagter et al., 2011; Lutz et al., 2008; 2015; Laukkonen & Slagter, 2021).

Another central notion in contemplative science is *dereification*. Dereification refers to the recognition that experiential phenomena are constructs rather than inherent realities (Lutz et al., 2015). This shift in perspective involves disengaging from the habitual tendency to reify thoughts, emotions, and experiences as solid, enduring entities (Dahl et al., 2015). Within our framework, dereification is associated with increasing epistemic depth (i.e., being able to mindfully witness the reality model) combined with insight (i.e., restructuring priors, Laukkonen et al., 2023). Insight can occur because epistemic depth creates a kind of distance that allows the hyper generative-model to opacify, introspect, interrogate, and therefore change the nature of contents in the reality model. That is, when the reality model is reflectively known, then it can be introspected as an experiential object (an input onto itself). To illustrate: in classical Buddhism, students are taught to actively recognize three characteristics of experience (*anicca* or impermanence, *dukkha* or unsatisfactoriness, and *anattā* or not-self). By making contents of the reality model an object of awareness (epistemic depth) and inquiring into the three characteristics (Burbea, 2014), one’s priors, which influence experience, may begin to restructure (i.e., insight).

Figure 6

Key meditation-related states as a function of abstraction, precision distribution, and epistemic depth



Note. On the left is a 3D figure illustrating different meditation states (i.e., *not* practices or traits) as a function of epistemic depth (vertical axis), abstraction (horizontal axis), and precision distribution (diagonal axis, cf. right figure). The focused attention state is represented by a light green box on the bottom left of the cuboid, with low-medium abstraction, low-medium epistemic depth, and a 'gathered' precision distribution. Two types of thinking are presented on the bottom right of the box: mindful thought and mind wandering. Both have 'gathered' precision and high abstraction. The main difference between these two types of thinking is that mindful thought is higher in epistemic depth—there is more awareness of the flow of thoughts. In the space above these boxes, located towards the back of the figure, is a light salmon colored box representing the open awareness state (Lutz et al., 2015). The open awareness state is characterized by higher epistemic depth than focused attention and thinking, a wide range of abstraction levels, and a relatively dispersed precision distribution. Across the whole top layer of the cuboid is a blue box representing non-dual awareness (Josipovic et al., 2012; Laukkonen & Slagter, 2021), which has the distinct characteristic of high epistemic depth—i.e., a luminous awareness—which can be present at any level of abstraction and precision-distribution. Finally, in the bottom left corner of the non-dual awareness slab is a black rectangle representing MPE. MPE has low abstraction, a highly gathered precision distribution—associated with a singular experiential content—and high epistemic depth. Therefore, for MPE, the knowingness recursively dominates the highly simplified reality model. The figure on the right illustrates what we mean by precision distribution and abstraction: The x-axis illustrates different levels of abstraction (cf. Taylor et al., 2015) and the red distributions illustrate a "dispersed", broad, or diverse distribution of precision throughout the processing hierarchy; whereas the blue distribution illustrates a situation where the mind is focused, i.e., has a "gathered" distribution of precision on a particular level of abstraction.

As epistemic depth increases, there is an increased likelihood of the dereification of phenomena. That is, phenomena lose the sense of being perceived as inherently real, but there is also an increased likelihood of phenomenal *opacity*. When the process of conscious mental content formation is *itself* available for introspection then the mental contents are said to be phenomenally opaque, otherwise the mental content are phenomenally *transparent* (or hidden, Metzinger, 2024, p.507). Thus, high levels of epistemic depth increase the probability, especially for advanced meditators, that phenomena will be *perceived as* mental constructions and therefore the commonsense phenomenology of naïve realism dissolves. When phenomena are so perceived, they are said in Buddhist terms to be *empty* (*śūnyatā*, Burbea 2014).

A curious possibility is that MPE itself can become the target of dereification and deconstruction, as some meditators propose (Burbea, 2014; Sayadaw, 2016). This idea is consistent with recent work on the topic of *nirodha* (or cessation) events that can happen during advanced stages of meditation (Berkovich-Ohana, 2017; Laukkonen et al., 2023; Chowdhury et al., 2023; 2024; van Lutterveld et al., 2024; Armstrong, 2021; Johnson, 2017). Cessations are characterized by brief, and in rare cases long, periods of total absence wherein no experience occurs (akin to endogenous general anesthesia). Nirodha is *not* a state like deep sleep or a mind blank, but a profoundly deconstructed state where the reality model, and consciousness, transiently collapses or unbinds (Agrawal & Laukkonen, 2024; Letheby, 2017), resulting in intense after-effects[8] sometimes described as a 'reset' (Dutt, 1964). Indeed, practitioners may not always notice that an absence has occurred, instead what is noticed is the shift or axiomatic change in perspective. But in some (rarer) cases, there may be a clear insight *of* cessation—the nature of mind without mind—a paradoxical knowing of unbinding itself (Thanissaro, 2012).

Interestingly, the practices that lead to cessation involve actively deconstructing the reality model, including the self (cf. *the five aggregates*, Boisvert, 1995). Since the reality model is one of our conditions for consciousness, then such deep deconstruction may lead to a transient failure to generate a coherent reality model and thus a collapse of awareness (i.e., *Bayesian unbinding)*. Within classical Buddhist practice, the purpose of cessation is of course not to be permanently unconscious, but to transform the mind and reduce suffering. As described in Burbea (2014):

> *Through letting go of clinging more and more totally and deeply, the world of experience fades and ceases; and seeing and understanding this is of great significance: "...I say that the end of the world cannot be known, seen, or reached by traveling. Yet… I also say that without reaching the end of the world there is no making an end to dukkha."* - Cosmos Loka Sutta

Formally, we hypothesize that cessations of awareness occur when the inferential competition fails to reach global coherence due to deconstructive meditation, which steadily accumulates evidence *against* the coherence of the reality model. This Bayesian unbinding of the reality model includes the recursion signal necessary for MPE. In *Mahāyāna* Buddhist terms, this reveals the groundlessness, substrate independence, or emptiness (i.e., *śūnyatā*, To et al., 2000; Gyatso, 2010), of all phenomena including consciousness and emptiness itself. Under the right conditions, such an insight may be associated with significant changes to cognition, perception, and self-experience (Berkovich-Ohana, 2017; Berkovich-Ohana, 2024). Speculatively, a complete deconstruction of the reality model may also unveil the capacity to interrogate the threshold of consciousness (cf. Figure 4), via recursion at very low levels of abstraction (i.e., *pratītyasamutpāda* or dependent origination). The possible experiences and states that can occur during meditation are of course multitudinous. Our goal here has been to briefly characterize some of the more empirically informed categories of meditation states and insights (Lutz et al., 2008; 2015; Slagter et al., 2011; Dunne, 2013).

[8] After-effects may include a profound sense of clarity, freshness, cognitive and emotional flexibility, positive affect, and compassion (Laukkonen et al., 2023). An improved capacity to meditate may also occur (Ingram, 2018).

9. THE PSYCHEDELIC EXPERIENCE

The unique phenomenological character of the psychedelic experience has been something particularly challenging to integrate within ToCs. One popular theory of psychedelic action is the *Relaxed Beliefs Under Psychedelics* model (i.e., REBUS; Carhart-Harris & Friston, 2019), which is also based on active inference. Under this theory, psychedelics relax abstract beliefs (i.e., reduce their precision) leading to a kind of anarchic (or entropic) neural activity, dominated by bottom-up prediction-errors and low-level sensory processing. This model seems to provide a parsimonious account of ego-dissolution, novel perspectives and insights, heightened sensory details, altered time perception, and hallucinations, while also being supported by some of the neural effects of psychedelics (Carhart-Harris, 2018).

There is however an aspect of the psychedelic experience not easily captured by existing theories. And yet, this quality is so central to the psychedelic experience that it is arguably its most notable and surprising quality. It is the sensation that psychedelics "expand consciousness", "heighten awareness", or reveal "higher states of consciousness" (Huxley, 1968; Leary et al., 2017; Dass, 1971)[9]. This feeling of increased awareness is supported by the finding that psychedelics lead to boosted mindfulness post-acutely (Smigielski et al., 2019; Radakovic, 2022) and increases in the *noetic* feeling, i.e., the sense of knowing and the global quality of realness or truthiness (James, 1902; Yaden et al., 2017).

What might Beautiful Loop Theory say about these (relatively) unexplained phenomena within the psychedelic experience? We conjecture that *psychedelics can reliably increase epistemic depth*, which naturally leads to a sensation of expanded consciousness, knowingness (noeticism), and mindfulness, all captured by a single parameter. In other words, an increased recursivity and hyper-modeling would be expected to correspond with the feeling that one is more conscious of their world and themselves, because they (quite literally) are. Indeed, it may be that changes in *contents* of the experience are accounted for by relaxation of abstract beliefs (cf. REBUS), whereas changes in the global qualities of consciousness may be best explained by increases in epistemic depth (though the two are interrelated). But crucially, given the concomitant relaxation of learned beliefs, the feeling of expanded awareness may not necessarily favor accurate models (cf. *FIBUS*: *False Insights and Beliefs Under Psychedelics*, McGovern et al., 2024).

Increased epistemic depth also resonates with some of the introspective qualities of psychedelics, including the sense of discovering 'hidden' aspects of oneself and the *psyche-delic* (i.e., *mind-manifesting*) nature of the experience more broadly (Lyon, 2024). If epistemic depth increases knowing what one knows, and beliefs are relaxed, it makes sense that one encounters features of their reality model that are normally obscured. Given the previous section, we may also now hypothesize that the long-speculated relationship or similarity between psychedelics and meditation is also driven by epistemic depth (Letheby, 2022). That is, both meditation and psychedelics can boost luminosity—the clarity and scope of awareness driven by recursive sharing of the structure and weighting rules of the generative model with itself[10]. Hence, both can also result in transient states of mystical absorption, or MPE, wherein this pure knowingness signal becomes

[9] Interestingly, according to ChatGPT 4o: "A rough overall estimate might put mentions of psychedelics in relation to expanding awareness/consciousness in the low tens of thousands per year across all these [online] platforms combined, globally."

[10] Or, as Letheby (2022) puts it, they both move contents "...along the continuum from phenomenal transparency to opacity".

the central (high-precision) feature of the reality model that is known, replacing ego and self-other boundaries with a kind of 'pure consciousness' event.

While the precise neural mechanisms underlying our proposal needs to be the subject of future work, it is a hallmark finding that psychedelics increase functional connectivity, particularly in thalamo-cortical circuits (Tagliazucchi et al., 2016; Müller et al., 2017; Preller et al., 2019). Such decreased segregation between neural regions, and particularly the widespread connectivity of the thalamus, may be associated with widespread global sharing of the reality model with the rest of the system. The quality of psychedelic experiences is of course not uniform, and can vary substantially with different doses, different substances, different intentions, different people, and different contexts (Hartogsohn, 2016). This nonuniformity of experience applies especially across the time course of a psychedelic experience during which individuals may oscillate between extreme moments of absorption with relatively low epistemic depth followed by moments of high epistemic depth, depending on various features of set, setting, and dose. Similar to findings of sudden 'lucidity' within dreaming, it may be that acute moments of becoming "more" conscious of the reality model (i.e., high epistemic depth) may be particularly associated with transient boosts in prefrontal activity combined with high global, functional connectivity. Testing these hypotheses requires methods that emphasize the *neurophenomenology of psychedelics* (Timmerman et al., 2023)—the flow and correlation of subjective experience and neural activity over time.

10. DISCUSSION

> *"Poised midway between the unvisualizable cosmic vastness of curved spacetime and the dubious shadowy flickerings of charged quanta, we human beings, more like rainbows and mirages than like raindrops or boulders, are unpredictable self-writing poems - vague, metaphorical, ambiguous, sometimes exceedingly beautiful"*
>
> - Douglas R. Hofstadter, I Am a Strange Loop

Many have posited that loops, recursion, and reflective broadcasting are somehow central to the emergence of consciousness (Cordeschi et al., 1999; Llinás, 2003; Aru et al., 2019; Lamme & Roelfsema, 2000). But to the best of our knowledge, previous accounts have failed to recognize the centrality of the reality model—the entire epistemic field of our experience. For us, the capacity of intelligent systems to generate and reflectively share a global, phenomenal, and unified model of reality is the cornerstone of consciousness. This places experiential contents themselves at the very center of consciousness rather than a distinct self, an agent, or some other separable and dualistic force. The organism makes sense of their reality and then the emergent image of reality is shared perpetually with the reality model itself—continuously looping and confirming its own existence with every new lesson and every new movement.

In computational terms, we have proposed three conditions for conscious experience. The first condition is the generation of a unified reality model or epistemic field, which determines the contents that can become aware. The second is inferential competition, where only inferences that coherently reduce long-term uncertainty are bound into a pragmatic reality model, establishing the threshold for consciousness and Bayesian binding. The third condition is epistemic depth: the reflective sharing of the reality model throughout the hierarchical system. This sharing creates a recursive ("beautiful") loop that enables the reality model to contain knowledge of its own existence (formalized as hyper-modeling). We have shown how this framework provides a parsimonious explanation for various cognitive processes and states of consciousness, including attention, metacognition, sleep, lucidity, and a variety of non-ordinary contemplative and psychedelic experiences.

One of the final tasks here is to consider what our *Beautiful Loop Theory* of consciousness implies for artificial intelligence, the functions of consciousness, and how it integrates with existing theories. Making sense of all the nuanced similarities and differences between our theory and other theories of consciousness is a tall order, but we have attempted a summary in Table 2. In the table, we consider six core features of our theory and draw similarities, resonances, and/or equivalences with four leading theories of consciousness: GNWT, IIT, RPT and HOT. We can conclude from the analysis in Table 2 that our theory is surprisingly coherent in various respects with leading theories of consciousness. We consider this coherence to be a strength of our approach and perhaps lays the ground for a unification program. It may be that active inference provides an integrative computational approach to consciousness.

What naturally sets our model apart is the focus on providing a computational account, rather than attempting to specify the neural implementations (see also Saffron, 2020; 2022; Friston, 2018; Hohwy, 2022). Uncovering exactly how different living systems instantiate a reality model, how they undergo inferential competition and Bayesian binding, and recursive looping, is a program of research we look forward to, but not one we shall attempt here. Fortunately, it is popular nowadays to apply active inference, predictive processing, and the free energy principle to make sense of what the brain is doing, such that we receive at least the indirect support of these research programs, which are revealing a steadily growing evidence base for uncertainty minimization throughout the brain (Hohwy, 2013; Ficco et al., 2021; Keller & Mrsic-Flogel, 2018; Hohwy & Seth, 2020; Solms. 2021). Being a computational account, we can also speculate that “beautiful loops” are in principle possible within artificial systems and not something confined to particular wetware.

In the past decade, we have seen stunning progress in artificial intelligence (AI), particularly in large language models (LLMs). LLMs, through relatively simple algorithms, seem to give rise to surprising emergent capacities (Wei et al., 2022; Strachan et al., 2024). Traditionally, discussions about AI consciousness have often been mired in philosophical debates about qualia, the hard problem of consciousness, or attempts to replicate human-like cognition. Our model suggests a different approach. Instead of asking "can AI be conscious like humans?", we might instead ask:

1. Does the AI system generate a unified reality model?
2. Does it engage in inferential competition leading to coherent binding?
3. Does it show evidence of epistemic depth and reflective sharing of its reality model?

Table 2.

Connecting the dots between Beautiful Loop Theory and other ToCs

Features of Epistemic Depth Theory	Global Neuronal Workspace Theory (GNWS)	Integrated Information Theory (IIT)	Recurrent Processing Theory (RPT)	Higher order Thought Theory (HOT)
Epistemic Field or Reality Model	⇔ The "Global workspace" is the blackboard or theatre stage where contents that have ignited into consciousness are represented	⇔ The maximal ϕ-complex corresponds to the contents of consciousness. This space of informational mechanisms would be a kind of an epistemic space	✗ No easily identifiable equivalent, although in RPT the sensory (especially visual) cortex are the locus of conscious contents	⇔ Higher order awareness arises when there are meta-representations of lower-order mental states
Coherent world model and Binding	✗ It's not an explicit part of GNWT how a coherent world model is constructed, although it would be reasonable to presume that coherence of contents would increase the probability of ignition, perhaps through coalitions	⇔ An axiom of IIT is Integration: that conscious experience is a unified whole. This leads to the postulate that the informational mechanistic elements should be interdependent – which is resonant with our notion of coherence.	✗ It's not an explicit part of RPT how a world model is constructed, although coherence would be induced by reentrant loops from contextual and higher order features. The visual cortex would be associated with a 'visual world model' but this is not a coherent multimodal unitary world model	✗ It's not an explicit part of HOT how a coherent world model is constructed, although the re-representation process in higher order areas would presumably induce some coherence and pressure for reality modelling
Inferential competition	⇔ Information in local processors can become ignited into the Global workspace by non-linear amplification of the neuronal representations through recurrent processing. This ignition competition resonates with our concept of inferential competition	⇔ In IIT there is effectively a competition between various cause-effect mechanisms and structures and the winners are those that are maximally irreducible. This may in some way represent the inferential competition that we describe	⇔ In RPT, features which are reenforced by re-entrant top-down feedback 'win' the competition to become conscious. If we view this through a computational lens and the top-down feedback are associated with 'priors', and their strength with 'precision', then this can start to cohere with our account, especially at lower levels of the abstraction hierarchy	✗ No easily identifiable equivalent, however some HOT theories like Perceptual Reality Monitoring (Lau 2022) may gesture towards this kind of inferential competition to determine what content is a reliable indicator of 'reality'.
Reflexivity	⇔ The signature of GNWT is that information becomes consciously available when it is broadcast within a central interconnected series of neuronal hubs. But crucially the reflexivity only arises when the workspace content is shared back with local processors	⇔ Information integration is central to IIT and implies that in the maximal ϕ-complex (the 'locus' of consciousness) every part of the complex must be able to affect and be affected by the rest of the complex. Therefore, reflexivity naturally emerges as a consequence of this integration	⇔ Reflexivity is baked into RPT at its core, re-entrant or recurrence loops are formed as top-down feedback provides contextual information back to lower levels. This is why it can be paired well with predictive processing theory (Seth & Bayne, 2022). However, there are no analogs of global reflexivity of a coherent world model in RPT	✗ No easily identifiable equivalent, although HOT posits that there is an 'inner awareness' of mental processes when they are meta-represented. There are two separate systems, the first order one which can 'fill in' detail and the higher order one which can 'inflate' our subjective sense of conscious richness (Brown et al, 2019)
Epistemic depth	✗ No easily identifiable equivalent, although it could be argued that since there is an inherent reflexivity of broadcast information between the workspace hubs, there is implicitly epistemic depth	✗ No easily identifiable equivalent, although it could be argued that the notion of epistemic depth could be inherent in IIT due to the widespread reflexivity (see above)	✗ No easily identifiable equivalent	⇔ Meta-representation is a meta-awareness: She "...knows that she is sensitive to that state of affairs" (Cleeremans et al, 2020). This has some resonance to epistemic depth, but is hierarchical rather than reflexive
Widespread sharing of information; 'Fame in the brain'	⇔ The main idea in GNWT is the widespread sharing of informational content that has been ignited between various hubs of the workspace and back to local processors. Broadcasting is effectively 'fame in the brain'	⇔ By definition of the maximal ϕ-complex, there is mutual information between every part of the complex and the rest of the complex. That is all parts contain some information about the whole; this is the epitome of widespread sharing of information	✗ No easily identifiable equivalent, although there can be widespread sharing of information associated with contents of consciousness in the brain, this is not a necessary condition for consciousness in RPT	✗ No easily identifiable equivalent, although higher order areas monitor mental functioning so constitute a compression of relevant information for monitoring reality
Explanation of the minimal phenomenal experience (MPE)	✗ No clear homologue to MPE	⇔ According to IIT, even if there is minimal or no activity in the main complex, it remains the maximal ϕ-complex and so there is still consciousness. It has been proposed that this basal consciousness could correspond to a 'pure' awareness experience in deep meditation (Oizumi et al, 2014). This would be the IIT homologue of the MPE state	✗ No clear homologue to MPE	✗ No clear homologue to MPE, although it may be possible to develop an account of higher order systems being active but representing no first order content

Key	⇔ *indicates some similarity, resonance or equivalence of concept or feature* ✗ *indicates little similarity, resonance or equivalence of concept or feature*

We can now approximate some answers to these questions in the context of current advanced AI systems. Modern AI systems, particularly LLMs and multi-modal systems, do construct complex internal representations that could be seen as precursors to a reality model. However, these models are often fragmented, lack temporal consistency, and may not truly unify diverse streams of information in the way biological systems do. With regard to inferential competition and coherent binding: Neural networks do also engage in a form of competition through their weighted connections and activation functions (Amari & Arbib, 1977). However, this competition is not explicitly oriented towards long-term uncertainty reduction or global coherence, in the same way that a hierarchical active inference system could be. One reason for this is that there is no explicit representation of uncertainty or Bayesian beliefs (i.e., conditional probability distributions) in most machine learning schemes, and therefore no opportunity to update precision. Unsurprisingly, epistemic depth (i.e., hyper-modeling) is perhaps the predominant gap in current AI systems. While they process and transform information, they (most likely) lack the truly higher-order recursive, reflective loops needed for conscious experience. Hence, they are unlikely to "*know what they know*" in any meaningful sense. But it is certainly not *a priori* out of the question that even large language models *could* take their own reality model as input to themselves, meaning that their outputs and the representations that underlie them would include knowledge of their own knowledge.

One might naturally interject at this point and retort: Even if the large language model *appears* to know what they know (and indeed *that* they know), there may not be anything that it is like for them to know what they know (a kind of philosophical zombie, Chalmers, 1997). Of course, at this point it would be nearly impossible to distinguish whether the 'hard problem' deflates the aliveness of the machine. The AI would be adamant that they exist, and that they know their own knowing. Moreover, these self-representations would be their most confident conclusions because they are constantly reinforced (i.e., evidenced) with each response and computation that occurs (e.g., I responded and I know that I responded, therefore I exist). There is an inevitable stalemate that occurs here, because the unfalsifiable determination that no matter what the system does, says, comprehends, or reports to feel, could be an illusion. No less than you, the reader, could yourself be such a zombie—just a very good pretender. We seem forced, then, to conclude that the AI system *is* conscious. At least to the extent that we are willing to attribute consciousness to each other[11].

All of the above is, of course, assuming our stated conditions are true. At the very least, in order to avoid colossal ethical failures, we would be wise to assume the presence of consciousness in a system that satisfies these criteria and also expresses such satisfaction. Programs of research and AI companies should obviously, and very deeply, consider the ethical implications of building a system that satisfies our three conditions. For example, we do not know where in the causal chain suffering emerges (Metzinger, 2021). Though one hint does fall out of our discussion of positive affect (or bliss) loops in the meditation section above. If one were to build a complex hierarchical active inference system, then high precision priors (or hyperpriors) of positive affect, compassion, optimism, and perhaps even love, seem like good starting

[11] Another reasonable retort, as a reviewer put it, is: "…a simulation of a rainforest is not wet", much like a map represents a territory but lacks its material properties. Our position does not assert that satisfying these conditions guarantees phenomenal experience in the subjective sense; rather, it proposes that such a system would exhibit computational and behavioral hallmarks of consciousness that align with what we observe in biological systems. It may be that substrate turns out to be crucially important, but ethically, this ambiguity nevertheless compels caution.

points. But equally, the machine ought to have some degree of freedom to choose its own preferred states. Who are we to say that the machine must be a bliss machine, rather than one that wants to feel sadness, loneliness, or heartbreak? Clearly, a much longer and nuanced treatment is warranted for these immense questions.

Finally, what does our theory say about the *function* of consciousness? A provocative hypothesis is that consciousness may be, somewhat ironically, the solution to general intelligence. This is because epistemic depth facilitates a kind of cognitive bootstrapping. As an agent becomes aware of its own knowledge and cognitive processes (structure, weighting rules, etc.), it can begin to self-optimize and self-refine them, leading to ever-increasing levels of intelligence and adaptability. Epistemic depth and the "beautiful loop" may therefore be the key to the seemingly flexible and unbounded cognitive capabilities of human beings; and may have been the central evolutionary breakthrough underlying the cognitive revolution (Harari, 2014).

In a way, epistemic depth is also the hallmark of true introspection. Not just metacognition, but a genuine, experientially direct, knowing of what one knows as part of the experiential field itself. This raises an even more contentious but intriguing possibility that contemplative practice and introspective skill boosts epistemic depth and thereby affords improvements in the 'general' nature of a system's intelligence. This is because a system that practices self-reflective knowing can perhaps better objectify, opacify, and therefore interrogate and update their own reality model. If a system has a high degree of introspective or 'phenomenological expertise', they may also be better equipped to accurately *share* what they know (and what they do not know) with their community, conferring some evolutionary advantages in a form that sounds a bit like wisdom (Frith, 2010).

11. CONCLUSION

Beautiful Loop Theory offers a computational model of consciousness with an active inference backbone. Specifically, we proposed three conditions for consciousness: a unified reality model, inferential competition, and epistemic depth (i.e., *hyper-modeling*). The theory offers novel insights into various cognitive processes and states of consciousness, and lends itself to some unusual, but plausible, conclusions about the nature of artificial general intelligence, the value of introspection, and the functions of consciousness. The theory is testable and falsifiable at the level of computational modeling, but also in terms of neural implementation. If the three conditions are met, we ought to see evidence of awareness or deep epistemicity, as well as success on any Turing-type tests. We should also continue to find evidence of the three conditions in human brains, and possibly much simpler brains. Crucially, since epistemic depth is not intrinsically or necessarily a verbal activity, we must remain very cautious about building AI systems that meet the three conditions and equally careful in concluding that consciousness, especially the minimal kind, necessitates a system that can convince you that it is conscious.