Towards a formal neurophenomenology of metacognition: modelling meta-awareness, mental action, and attentional control with deep active inference

Preprint · June 2020

DOI: 10.31234/osf.io/5jh3c

CITATIONS
11

READS
978

6 authors, including:

Lars Sandved-Smith
French Institute of Health and Medical Research
7 PUBLICATIONS **85** CITATIONS
SEE PROFILE

Casper Hesp
University of Amsterdam
57 PUBLICATIONS **988** CITATIONS
SEE PROFILE

Antoine Lutz
French Institute of Health and Medical Research
143 PUBLICATIONS **16,695** CITATIONS
SEE PROFILE

Jérémie Mattout
French Institute of Health and Medical Research
163 PUBLICATIONS **7,428** CITATIONS
SEE PROFILE

Title:
Towards a formal neurophenomenology of metacognition: modelling meta-awareness, mental action, and attentional control with deep active inference

Authors:
Lars Sandved Smith[1,2] (lars.sandvedsmith@gmail.com)
Casper Hesp[2,3,4,5] (c.hesp@uva.nl)
Antoine Lutz[1] (antoine.lutz@inserm.fr)
Jérémie Mattout[1] (jeremie.mattout@inserm.fr)
Karl Friston[2] (k.friston@ucl.ac.uk)
Maxwell J. D. Ramstead[2,6,7] (maxwell.ramstead@mcgill.ca)

Affiliations:
1. Lyon Neuroscience Research Centre, INSERM U1028, CNRS UMR5292, Lyon 1 University, Lyon, France.
2. Wellcome Centre for Human Neuroimaging, University College London, London, UK, WC1N3BG.
3. Department of Developmental Psychology, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands.
4. Amsterdam Brain and Cognition Centre, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands.
5. Institute for Advanced Study, University of Amsterdam, Oude Turfmarkt 147, 1012 GC Amsterdam, Netherlands.
6. Division of Social and Transcultural Psychiatry, Department of Psychiatry, McGill University, Montreal, Quebec, Canada.
7. Culture, Mind, and Brain Program, McGill University, Montreal, Quebec, Canada.

Abstract:
Metacognition refers to the capacity to access, monitor, and control aspects of one's mental operations and is central to the human condition and experience. Disorders of metacognition are a hallmark of many psychiatric conditions and the training of metacognitive skills is central in education and in many psychotherapies. This paper provides first steps towards the development of a formal neurophenomenology of metacognition. To do so, we leverage the tools of the active inference framework, extending a previous computational model of implicit metacognition by adding a hierarchical level to model explicit (conscious) meta-awareness and the voluntary control of attention through covert action. Using the example of mind-wandering and its regulation in focused attention, we provide a computational proof of principle for an inferential architecture apt to enable the emergence of central components of metacognition: namely, the ability to access, monitor, and control cognitive states.

Keywords: Attention, metacognition, opacity and transparency, mental action, active inference, free energy principle, focused attention, neurophenomenology

Acknowledgements:
We are grateful to Laurence Kirmayer, Soham Rej, Bassam El-Khoury, Andy Clark, Mark Miller, Jakub Limanowski, and Michael Lifshitz for helpful comments and discussions that helped to shape the content of this paper. This research was supported by a Research Talent Grant of the Netherlands Organisation for Scientific Research (no. 406.18.535)(CH), the LABEX CORTEX of Université de Lyon (ANR-11-LABX-0042) within the "Investissements d'Avenir" program (ANR-11-IDEX-0007) (LSS,AL,JM), by a European Research Council grant (ERC-Consolidator 617739-BRAINandMINDFULNESS) (AL), by a grant from the French National Research Agency ("MindMadeClear," ANR-17-CE40-0005-02) (AL,JM), by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z) (KJF), and by the Social Sciences and Humanities Research Council of Canada (MJDR).

1. Introduction

1.1. Towards a scientific study of metacognition

Metacognition – cognition about cognition – is a hallmark of human phenomenology. The mechanisms, effects, and dysfunctions of such processes are highly relevant to cognitive, clinical, and theoretical neurosciences. The metacognitive control of attentional states, for instance, is thought to be at the heart of several psychiatric conditions. Recent work suggests that this is the case in schizophrenia (Brown et al., 2013) and in autism (Kiverstein et al., 2020; Van de Cruys et al., 2014). Although they are not often or primarily described as expressions of attentional deficits, the latter two conditions have been related to failures of metacognition. Both involve specific and subtle dysfunctions that affect perception, thought, and mood, but which stem from atypical inference about—and faulty control over—one's higher-order states and processes. These are of higher order in that they govern the attribution of confidence to one's own perceptions (Lysaker et al., 2020; Palmer et al., 2017).

In cognitive sciences, theories on metacognition (Fleming et al., 2012; Nelson, 1996; Schooler & Smallwood, 2009) typically distinguish between (1) metacognitive *knowledge*, that is, the information one agent can gather about how the mind works; (2) metacognitive *monitoring*, that is, the observation and following of one's mental operations; (3) and metacognitive *control*, that is, self-regulatory actions made possible by the metacognitive monitoring. Metacognitive monitoring can be performed with or without *explicit* awareness (Reder & Schunn, 2014), highlighting a distinction between non-conscious tacit monitoring and conscious *meta-awareness* (Schooler & Smallwood, 2009). Meta-awareness is defined in this context as the ability to *explicitly* notice the current content of consciousness. A canonical example of meta-awareness is to become aware of an episode of mind-wandering (Schooler et al., 2011). Metacognitive action, in turn, refers to overt behavioural adjustments as well as mental actions, such as covert attentional control, during a visual task. Generally speaking, metacognition is useful because it allows agents to increase their control over themselves by, e.g., allowing them to revise their beliefs without having to sample new data.

The aims of this paper are threefold: (1) to explicitly model meta-awareness, (2) to account for the computational consequences of meta-awareness on metacognitive control, and (3) to provide a computational account of mental (covert) action. Formally speaking, the crucial thing to note about these cognitive processes—that underwrite metacognition—is that they are *about* other such processes. The deployment and control of attentional states and processes, for instance, is quintessentially a metacognitive ability: such states and processes monitor and modulate other, typically lower-level, perceptual states. Higher-level processes, in turn, oversee attentional processes: we can be aware of our attention being grabbed and make a deliberate decision to focus our attention elsewhere. It would seem, then, that effective self-regulation of attention depends on the ability to access, evaluate, and control the quality of these attentional processes themselves; in the same way that attentional processes are necessary to consciously access, assess, and control lower-level perceptual states.

In computational terms, the central place of confidence and reliability in effective self-regulation speaks to the key role of the *precision* (or inverse variance) of implicit beliefs[1] and the way they are estimated, optimised, and controlled (Hesp et al., 2020). Precisions, in a nutshell, quantify our confidence in our beliefs; say, how confident we are about what we know about states of the world, about their relation to our sensory observations, or about how these states change over time. By analogy with physical action, a *mental* (or *covert*) action consists in deploying and adjusting these key quantities, without there necessarily being an explicit behavioural counterpart to these covert actions (Limanowski & Friston, 2018, 2020). The nascent field of computational psychiatry is largely about deciphering these precision-related processes, shedding light on their physiological implementation (Lecaignard et al., 2020) and establishing their causal link with specific clinical traits (Friston et al., 2017).

Covert or mental actions, as just defined, also speak to other fields of neuroscience. In some traditions, such as mindfulness meditation, an objective of training is to make cognitive processes accessible, and hence controllable. The field of meditation research, sometimes referred to as contemplative neuroscience, has grown rapidly since the early 2000s (Eberth & Sedlmeier, 2012; Fox et al., 2016; Sedlmeier et al., 2012; Tang et al., 2015). This literature has increased our understanding of the relationship between meta-awareness and metacognition, especially with regards to attentional processes. Mechanistic models of these processes are beginning to appear (Farb et al., 2015; Jamieson, 2016; Lutz et al., 2019; Manjaly & Iglesias, 2019; Pagnoni, 2019). Complementary to this work, the contribution of this paper is a formal and computational architecture of these processes derived from first (Bayesian or variational) principles, which explicitly disambiguates the relationship between meta-awareness and metacognition.

Several conceptual frameworks exist in phenomenology, clinical psychology, and cognitive sciences to describe this relationship. For the purpose of this work, we focus on *regulatory* metacognitive control strategies, in which an agent seeks to control its mental states by becoming aware of the state as a cognitive process (as opposed to regulation strategies where one remains engrossed in the contents of the state). This style of regulation is central in mindfulness-related intervention, where patients learn to change their relation to thoughts and emotions, rather than change the thoughts and emotions themselves. As such, this regulation accounts for the positive effect of mindfulness meditation on mood disorders (Segal & Teasdale, 2018; Wetherell et al., 2017). The ensuing metacognitive perspective on mental states has been labelled as phenomenological reduction (Varela, 1996), decentering (Bernstein et al., 2015), cognitive defusion (Fletcher & Hayes, 2005), mindful attention (Papies et al., 2012), dereification (Lutz et al., 2015), or opacification (Metzinger, 2003). In contrast to this stance, being self-immersed in the contents of one's mind has been called cognitive fusion (Fletcher & Hayes, 2005), reification (Lutz et al., 2015), absorption

[1] That is, probabilistic or posterior Bayesian beliefs that describe a probability distribution over some latent quantity. These beliefs are subpersonal; however, the argument pursued in this work is that Bayesian beliefs about Bayesian beliefs can, in certain situations, become propositional.

(Tellegen & Atkinson, 1974), experiential fusion (Dahl et al., 2015), subjective realism (Lebois et al., 2015), or transparency (Metzinger, 2003).

To operationalise this aspect of metacognition, we will follow the distinction found in the self-model theory of subjectivity (Metzinger, 2004), as it has already been used in previous treatments which directly influence the present work (Limanowski & Friston, 2018); namely, the distinction between opacity and transparency.

1.2. Target phenomenology: Opacity and transparency

The capacity of a system to access some subset of its own states has been theorized under the rubric of 'opacity' versus 'transparency' (Metzinger, 2003). According to this framework, the mental states of human beings can be broken down into two kinds: those that are accessible to the system per se, which are labelled as 'opaque' (in the sense of being perceptible), and those that are not, and are labelled as 'transparent' (in the sense of being imperceptible). Some mental processes function only to make aspects of the world perceivable. We are not aware of them *as such*, but rather, we are aware of the content that they make available: these cognitive processes are 'transparent,' like a glass window that allows us to see what is outside. Other processes, however, make these cognitive constructive processes accessible per se. This second set of processes are metacognitive processes, about other states of the mind, to which they provide access, as a new source of data now made available for further processing. These processes are akin to the scroll wheel on a pair of binoculars, which has a position state that its user can control, and which enables one to apprehend, and to control the precision of, sensory inputs.

Transparent states and processes are thought to be by far the more common kind, which we share with most other animals. They mediate the agent's access to the latent causes of its sensory states, to things appearing "out there" in the world. The basic idea, then, is that we are not conscious of many of our mental states (the transparent ones) *as being* mental states, i.e., grasping them explicitly as being the results of constructive cognitive processes. Rather, such processes allow us to access some content, which is not experienced as constructed or mental. Accordingly, Metzinger argues that a state is transparent just in case the processes that construct it or constitute it are *not available* to the system through *introspective attention* (Metzinger, 2003). For example, the process of dreaming is a transparent state until the dreamer becomes lucid, i.e. aware of the fact they are dreaming, at which point the dreaming process becomes opaque (i.e., perceptible as a process that can be controlled).

Generalising from the concept as it figures in the self-model theory of subjectivity, we can say that some set of states and processes is partially *opaque* when the cognitive agent (1) usually employs them transparently, to interact with things, events, and places appearing subjectively real in the world or the mind, but (2) is also able to represent these states to itself or *access* them as well, as data for further inference. They are fully *transparent* when condition (2) does not hold. The processes labelled 'metacognitive' are related to the hallmark intentional features of opacity, in that they render other states opaque. (Note that

metacognitive states themselves are not *necessarily* opaque; as we will see below, to make them accessible requires a further set of processes.) Attentional states, for instance, are *about* perceptual states – they are second-order states; just as precisions are second-order parameters pertaining to first-order parameters (their mean values). In other words, they are about the results of an inferential process (Parr & Friston, 2017). Or again, consider emotional states and processes, which are about interoceptive and exteroceptive states (Allen et al., 2019; Hesp et al., 2020). In all these cases, the states, and processes at play in attention and emotion not only guide behaviour implicitly; they can be grasped as such, as when we introspectively reflect. This architecture of metacognitive opacity is thought to be that which underwrites metacognitive capacities in general.

This paper takes steps towards the development of a formal, computational account of metacognition as defined above in terms of meta-awareness and metacognitive control (e.g. the control of attention) through the deployment of mental actions. This will take the form of a *formal neurophenomenology* of metacognition. Formal neurophenomenology is an influential approach to the naturalistic study of conscious experience (Lutz, 2002; Ramstead, 2015; Varela, 1997). The aim of this approach is to formalise the aspects of lived experience that are revealed by phenomenological description – e.g., classical phenomenological accounts of the lived experience of moving about as an embodied agent (Merleau-Ponty, 1945) or having an inner consciousness of time (Husserl, 1927) – by providing a model of the inferential processes that would have to be in play for an agent to have that kind of experience, allowing for the target phenomenology to be experienced as such (for such an account of inner time consciousness, see (Grush, 2005; Varela, 1997; Wiese, 2017)).

1.3. Computational modelling of metacognition

We build on recent advances in computational modelling that are shedding light on the inferential processes underpinning the perception and behaviour of embodied organisms. These technical advances make it possible to implement self-reflective, hierarchically structured inferences and simulate behavioural dynamics that can be formally linked to metacognition and the execution of covert (mental) actions, such as attentional control.

At the root of these advances is a biologically plausible, neurocognitive and behavioural modelling framework called *active inference* (Friston, 2019; Friston et al., 2016). Active inference provides us with a Bayesian mechanics: that is, a mechanics of knowledge-driven action and active perception, which explains from first principles how autonomous agents can thrive within their ever-changing environments. Active inference descends from, and is closely related to, older and more familiar Bayesian theories of the brain, such as the Bayesian brain hypothesis (Knill & Pouget, 2004) and predictive coding (Bastos et al., 2012; Rao & Ballard, 1999). It casts perception, learning, and action as essentially being in the same game: that of gathering evidence for the model that underwrites the existence of the agent (Friston, 2013; 2019; Ramstead et al., 2018). In this sense, active inference casts living and cognitive processes as self-fulfilling prophecies, which gather evidence for an implicit (generative) model that the agent embodies and enacts (Ramstead, Kirchhoff, et al., 2019).

Formal approaches to the study of the capacities for metacognition and meta-awareness have recently been developed that leverage *parametrically deep active inference* (Hesp et al., 2020). Parametric depth is a property of cognitive architectures, such that this architecture comprises nested belief or knowledge structures: that is, beliefs about beliefs and inferential processes that operate on (or take as their input) the results of other inferential processes, as data for further processing. By construction, such cognitive architectures are capable of a rudimentary form of access to self-states and metacognition. This is because having an internal structure (a generative model) that evinces parametric depth endows the agent with higher-level states that renders the results of lower-level inference (e.g., posterior state and precision estimates) available as data for further self-inference. This enables the agent to access and control crucial aspects of themselves (Limanowski & Friston, 2018).

The active inference model as it is currently formulated is limited to *implicit* metacognition; e.g., it is able to model the deployment of attentional and emotional *processes* that direct the flow of perceptual inference and action (Parr & Friston, 2017), but cannot yet model the agent's access to, and evaluation and control of, its own attentional *states*. Here, we extend active inference beyond this limitation. To do this, this paper extends the deep active inference framework to account for the agent's capacity for *explicit* meta-awareness. This extension endows the agent's generative model with a deep, tertiary, hierarchical architecture (see Methods): policy selection (i.e., the formation of self-fulfilling beliefs about action) can be formulated hierarchically, allowing us to construct an active inference model of an agent's awareness of, and control over, aspects of their own generative model (i.e., rendering these states accessible or opaque, and thereby making them controllable).

In the example of sustained attentional control, it becomes crucial to be aware of where one's attention is focused and to recognise shifts in attention (i.e., distractions) quickly, to then recalibrate attentional processes accordingly. We demonstrate that, during a focused attention task, agents possessing a parametrically deep generative model exhibit the phenomenological cycles of focus and mind wandering that have long been associated with focused attention and mindfulness meditation practices (Lutz et al., 2008). Furthermore, this deep architecture enables such agents to report on their metacognitive observations, such as the extent to which they are aware of where their attention is focused. This extension to the active inference framework makes it possible to plausibly simulate, and make behavioural predictions about, the processes at play in meta-awareness and metacognition. More generally, it provides a biologically plausible understanding, derived from first principles, of the computational architecture required for the emergence of central aspects of metacognition; namely, meta-awareness and explicit metacognitive control of attention.

In summary, in this paper, we argue: (1) that metacognition is predicated on higher level access to cognitive states; (2) that the distinction between partially opaque and fully transparent states and processes can be formalised under deep active inference via a formal treatment of meta-awareness and its implications; and (3) that understanding this inferential architecture constitutes a first step towards a formal, computational neurophenomenological

account of metacognition in general. The remainder of this paper is structured as follows. After reviewing the active inference framework, which forms the methodological and theoretical backbone of our proposal, we turn to numerical proofs of principle. We close by discussing the implications of our model for the study of consciousness, for the study of attention under active inference, and for the progress of computational psychiatry and neuroscience.

2. Methods

The methodology we employ rests on a probabilistic graphical model (or generative model) that captures the inferential architecture required for the phenomenon of interest; in this case, the opacity-versus-transparency distinction and ensuing forms of metacognitive monitoring and control. Having specified this model, in the following section, we implement this model in computational simulations that reproduce the target behaviour. We illustrate meta-awareness and metacognitive control by examining the emergent dynamics of focused attention and mind wandering. We will see that simulated agents endowed with this form of inferential architecture organically reproduce the relevant phenomenology.

2.1. Introduction to the active inference formulation

The key technical innovation presented in this paper is an extension of the active inference framework derived from the work of Hesp and colleagues (2020). It builds on previous work suggesting that to deliberately attend to a set of states, and thereby render them opaque, corresponds to the top down deployment of second-order inference, i.e., inference about confidence or precision (Limanowski & Friston, 2018, 2020).

Active inference models cast perception, learning, and behaviour as governed by a single imperative: to minimise variational free energy (Friston, 2019; Ramstead et al., 2018). This variational free energy is an information theoretic construct that, in general, quantifies the divergence between observed and expected data, under a probabilistic model of the process that generated this data (the generative model). In a nutshell, it scores the probability of each model of how the data was generated by quantifying (technically, by providing an upper bound on, or approximation of) how much evidence for the model is provided by the data, or how much the variance in our data is explained by the model. This measure or score is the complement of variational free energy; and inference based on finding the model associated with the least free energy is known as variational inference. To minimise variational free energy, then, is equivalent to maximising Bayesian model evidence (Friston, 2019; Friston et al., 2010) or self-evidencing (Hohwy, 2016).

Variational inference was originally developed in statistical mechanics to convert intractable inference problems into easier optimisation problems, where the difficult problems associated with inferring the latent causes of data is converted into an easier optimisation problem: namely, selecting the model associated with the least free energy (Feynman, 1972). In the

current context, we treat the brain as an empirical scientist that is trying to make sense of her world, through careful sampling of sensory data (e.g., visual palpation) and then selecting the perceptual hypotheses that have the greatest evidence (Gregory, 1980).

2.2. Generative models

In this context, variational free energy quantifies the discrepancy between observed outcomes (i.e., the sensory states of an organism) and the outcomes that would be predicted under a statistical (generative) model of how her sensory data was produced, which is thought to be implemented in the networks of the brain.

Generative models are so called because they are models of the *generative process* 'out there in the world' that cause (or generate) our sensory data (Friston et al., 2016). The generative process is typically specified using equations of motion that capture the dynamics of how the environment generates observations. The active inference agents themselves, of course, do not have access to the generative process, and must deploy inference and action to guess-timate its structure; but in practice, when simulating these dynamics we usually write down the 'true' process, as we will below. The generative models considered here are called Markov decision processes (MDPs), a commonly used Bayesian scheme that applies to discrete state-spaces.

The most basic generative model, written to account for perception, is depicted in Figure 1. This elementary generative model quantifies the relation between observations (denoted o) and the latent or hidden states (s) that caused them. Here, this relation is captured by a likelihood mapping, denoted **A**, which encodes beliefs about how states of the world are related to the observations that they generate (i.e., 'assuming that some hidden state s is the case, what is the probability of observing o?'). Technically, this likelihood mapping is implemented as a matrix (**A**) specifying the probability of a particular observation, given a hidden state; formally, this is denoted $P(o \mid s)$. The initial state vector, **D**, in turn, specifies beliefs about the most likely state of the world independently of any observation, formally, $P(s)$: these are known as prior (Bayesian) beliefs.

In this context, variational inference moves from the quantities that the system can access – i.e., its observations, o, its prior beliefs, $P(s)$, and its beliefs about how its observations are caused by states of the world, $P(o \mid s)$ – to the quantity that it is trying to infer, namely, the most probable cause of its sensations, $P(s \mid o)$.

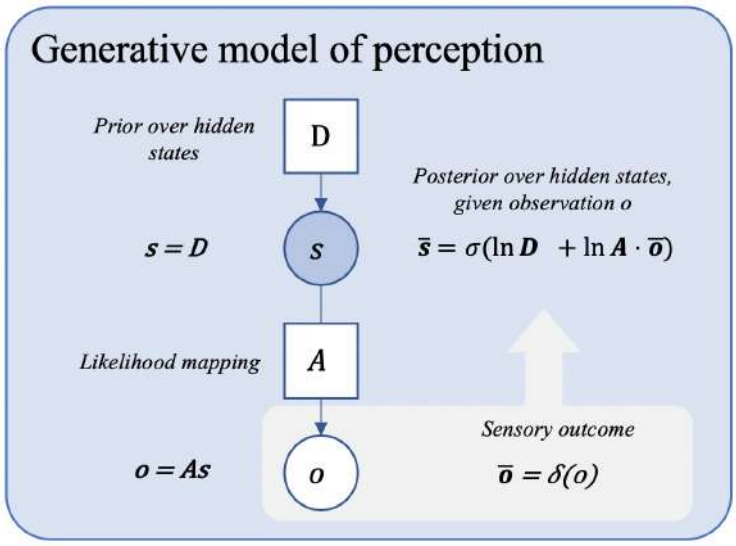

Figure 1. A probabilistic graphical model showing a basic generative model for moment-to-moment perception. This figure depicts a simple generative model for perception. Inference here inverts the likelihood mapping from causes to their outcomes, P(*o* | *s*), using prior beliefs about states (**D**) and sensory data (*o*), to obtain (or approximate) the most probable state P(*s* | *o*). Here, Bayesian beliefs are noted in bold, bar notation represents posterior beliefs, $\boldsymbol{\sigma}$ is the softmax function (returning a normalised probability distribution), and δ is the Kronecker delta function (returning 1 for the observed outcome, zeros for all non-observed outcomes). For the derivation of the latent state belief update equations shown see (Friston et al., 2016 Appendix A). The graphical presentation was adapted from a template of Figure 1a from Hesp et al. (2020).

We can always associate some level of confidence to each of the parameters described above. As discussed, precision is defined as the *inverse variance* of some distribution. Precision is also known as confidence and captures the degree to which the information encoded in the associated parameter is reliable. Of note is that precision already introduces some degree of parametric depth into the generative model because it is a second-order statistic: it is a belief about some other beliefs, namely, about how reliable they are.

In active inference, *attentional processes* have been formulated in terms of the precision (here denoted γ) of – or confidence in – the likelihood mapping (the **A** matrix; (Parr & Friston, 2017). Since they operate on the basis of second-order statistics (i.e., *precision*), attentional processes are seen as implicitly metacognitive in this framework. Intuitively, we can see why precision-modulation over **A** corresponds with attentional processes: the precision on **A** represents the extent to which the agent believes her observations *accurately map onto* hidden states. Attending to some stimuli increases the relative weight or gain on inferences made on the basis of that particular data or observation. For example, by paying closer attention to an ambiguous sound, the agent has greater confidence in determining the location of its origin than when the sound was first heard without being heeded. The process of combining available data with estimations of that data's reliability – in order to arbitrate its effect (relative to prior beliefs) on the overall inferential process – is known as *precision weighting* or *precision control.* Under active inference, this is the candidate mechanism for attentional modulation of perception (see Figure 2).

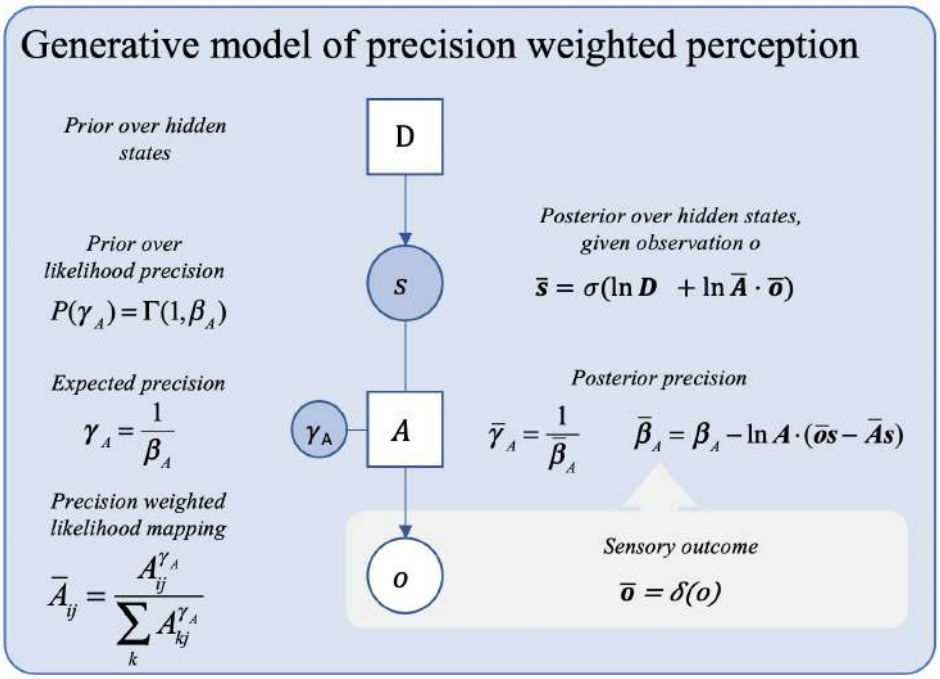

Figure 2. A Bayes graph showing a basic generative model of perception with precision. The precision term, γ_A, over the likelihood mapping **A** is sampled from a gamma distribution with inverse temperature parameter β_A. For the derivation of the precision belief update equation shown see (Friston & Parr, 2017 Appendix A.2).The graphical presentation was adapted from a template of Figure 1a from Hesp et al. (2020).

This basic generative model can only do posterior state estimation (i.e., perception) on a moment-to-moment basis: it does not encode knowledge that would allow its user to make predictions about the future. Active inference models, however, are not confined to processes unfolding from moment to moment. Indeed, these models have been extended to account for knowledge of temporal dynamics of states by inducing *beliefs about state transitions*, denoted **B**. These encode the probability of being in some state s_2 at some time step τ+1, given that the system was in state s_1 at time step τ, formally, $P(s_2 \mid s_1)$. Equipped with such Markovian models, an agent can effectively make inferences about future states of affairs.

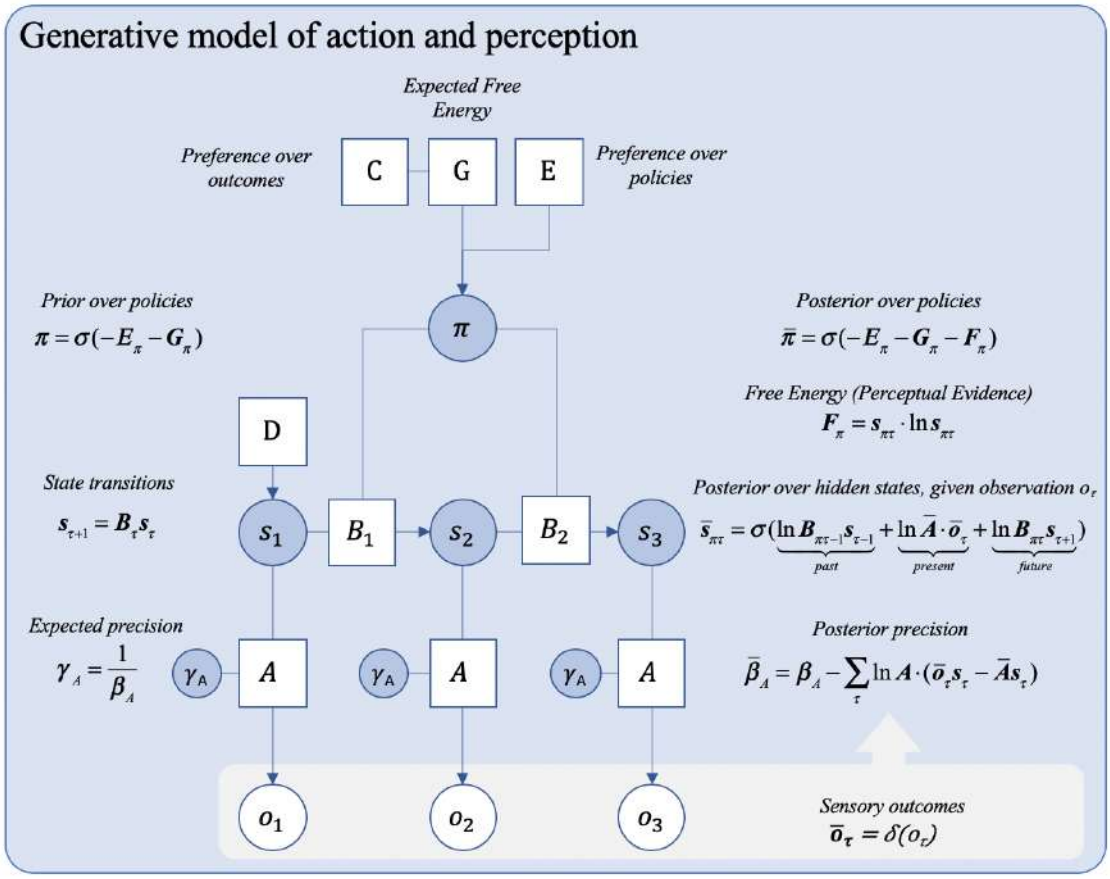

Figure 3. A Bayes graph showing a deep generative model for policy selection.
This model is equipped with beliefs about state transitions. Posterior state beliefs at each timestep now depend on beliefs about the previous and subsequent states, mediated by the state transition matrix **B**. Adapted from a template of Figure 2 from Hesp et al. (2020).

Equipping a model with beliefs about state transitions opens up a whole new domain of inference; namely, controlling observations through the selection of actions. As such, active inference is a game of *policy selection*. A policy, denoted π, is defined as a set of beliefs about which actions it is that one is undertaking. Policies are necessary because, in active inference, the agent must actively infer what course of action it is pursuing, on the assumption that what it is doing is likely to minimise variational free energy (Friston et al., 2016). Policy selection is implemented through the updating of beliefs about state transitions, now informed by the consequences of action. Intuitively, this implicit view of action is due to the fact that the agent does not have unmitigated or direct access to the actions that it undertakes. Rather, it can only access the sensory consequences of those actions; and so, it must infer what it is doing, given its action prior and its sensory data. A policy, then, is just a series of state transitions (**B** matrices) – and policy selection means anticipating that sequence of **B** matrices that is associated with the least expected free energy (**G**). Thus, as the name suggests, active inference casts action as a form of (variational) inference, as a self-fulfilling prophecy (Friston, 2011).

Of note is the prior preference mapping (or **C** matrix), which specifies prior beliefs about sensory outcomes, and the prior preference over policies (**E**). Both these beliefs affect policy selection directly. The **E** matrix encodes beliefs about what the agent would do, independent of the expected free energy in the current context. The expected free energy scores the posterior probability of different allowable policies in terms of outcomes. In the same way that maximising evidence minimises *inaccuracy* and *complexity*, selecting a policy that minimises expected free energy minimises *ambiguity* and *risk*. Here, risk is the difference between predicted and preferred outcomes under each policy. Preferred outcomes are parameterised by a **C** matrix that feeds into the calculation of the expected free energy for every policy, **G**: in short, prior beliefs implement a bias in the agent's action model towards policies that realise preferred outcomes.

2.3. Formalising metacognition, attention, and control under active inference

Finally, attentional, and metacognitive *states* can also be defined in terms of active inference (Hesp et al., 2020). In such a scheme, a further hierarchical level of states, and corresponding processes of posterior state and precision estimation, are introduced into the model. In this paper, we define two levels in addition to the first outlined above. The second level comprises *attentional states* that entrain likelihood mapping precisions at the first level. The third level comprises what one might call *meta-awareness states* (i.e., a state representing the degree to which one is aware of their own attentional state) that entrain likelihood mapping precisions at the second level. There is nothing special about these states; other than they generate outcomes at lower levels that speak – not to which state the world is in but – to the precision of beliefs about which state the world is in.

As we noted above, attentional processes always already involve some opacity, in the sense that they take as their input, and operate on, lower-level (perceptual) states. Deep active inference models add new hierarchical levels of state inference to exploit the parametric depth induced by lower-level state and (especially) precision estimation. In the parlance of the self-model theory of subjectivity (Metzinger, 2003, 2004), the use of precision and state estimation at lower levels as data or observations by higher layers of the model effectively renders the lower levels opaque. This involves conditioning precisions at the first level on *attentional control states* at the second level. Having defined attentional states, and having connected them appropriately to the precisions they control at the first level, we can treat them the exact same way that we treat other states; namely, by estimating their posterior expectation through (variational) inference. In effect, this means that our active inference agent has to infer in what attentional state it finds herself.

The next, crucial step is to define a transition matrix $\mathbf{B}^{(2)}$ at the second level, which specifies beliefs about *transitions between attentional states*. Since attentional states are just ordinary states – defined at the higher level – we can equip the model with state transitions at that level as well.

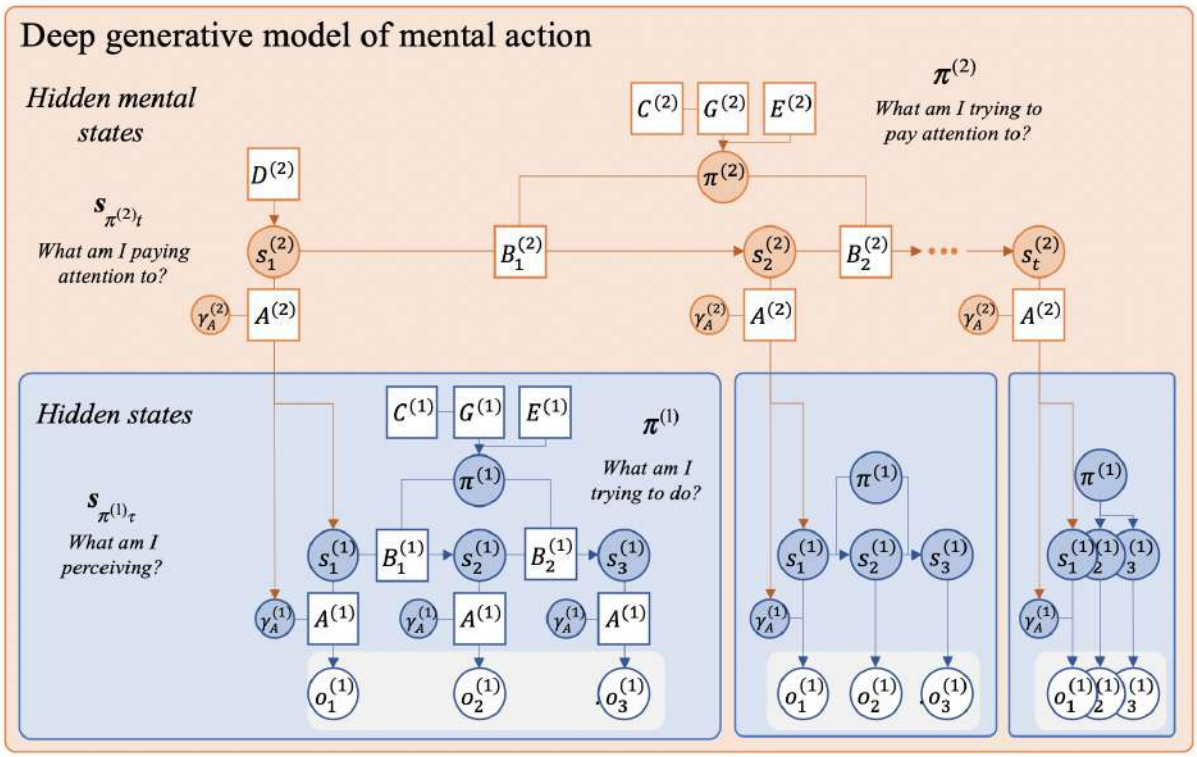

Figure 4. A probabilistic graphical model showing a deep generative model with second-order, attentional states. This deep generative model includes a new level of states, denoted $s^{(2)}$, which condition the precision on the first-order likelihood mapping $\mathbf{A}^{(1)}$. Attentional policies, $\boldsymbol{\pi}^{(2)}$, represent mental (covert) actions that condition transitions between attentional states. Adapted from a template of Figure 4 from Hesp et al. (2020).

Given this setup, we can define *mental (covert) policy selection*. Attentional states transition one into the other as well; and we have defined the $\mathbf{B}^{(2)}$ matrix at the second level to capture the agent's beliefs about those transitions. This setup means *attentional control* becomes *top-down state-dependent precision deployment*. Having defined mental state transitions, a mental policy can now be defined in the usual way: as a policy that conditions hidden state transitions at the second level, i.e., affecting the elements of $\mathbf{B}^{(2)}$. The key difference is that these hidden mental states themselves condition precisions at the lower level. This provides a formal treatment of mental action as deployment of precision as proposed by Limanowski and Friston (2018), by defining an appropriate generative model showing that covert action arises naturally from the formal definition of attention.

Finally, we can define a further level of state inference, which we can associate with the awareness of being in a given attentional state. These might be called *meta-awareness states*, which take as input the posterior state and precision estimates at the second (attentional) level. Recall that, to make perceptual states opaque on the first level, we defined a second level of the generative model, which takes as its data state and precision estimates at the first level. The same reasoning now applies to attentional states. In order to capture the phenomenology of deliberate, sustained attentional control, as found in meditation practices (Lutz et al., 2015), we define a third layer of hidden states, which represent latent meta-awareness states of the agent. Transitions between states at this level represent shifts in the

level of awareness that the agent has of where its attention is focused. The resulting three-level model of meta-awareness and metacognitive control is presented in Figure 5.

One might wonder whether such a level is necessary to model meta-awareness and opacity. We argue that it is. The simple reason is that such a level is necessary for *explicit awareness and control* of attentional states. Attentional states are hallmark metacognitive states; but for some traditions, such as mindfulness meditation, the question is less about the mechanisms that selectively enhance or suppress some aspects of experience, and more about accessing and assessing the quality of these attentional states and processes themselves. To deliberately control its attentional state, after all, the agent must be explicitly aware of it. This is to say that attentional states at the second level must also be made opaque, which is the work done by state inference at the meta-awareness level. In short, to be aware of anything (including awareness) one has to destroy its transparency by inferring and deploying precision control (i.e., mental action).

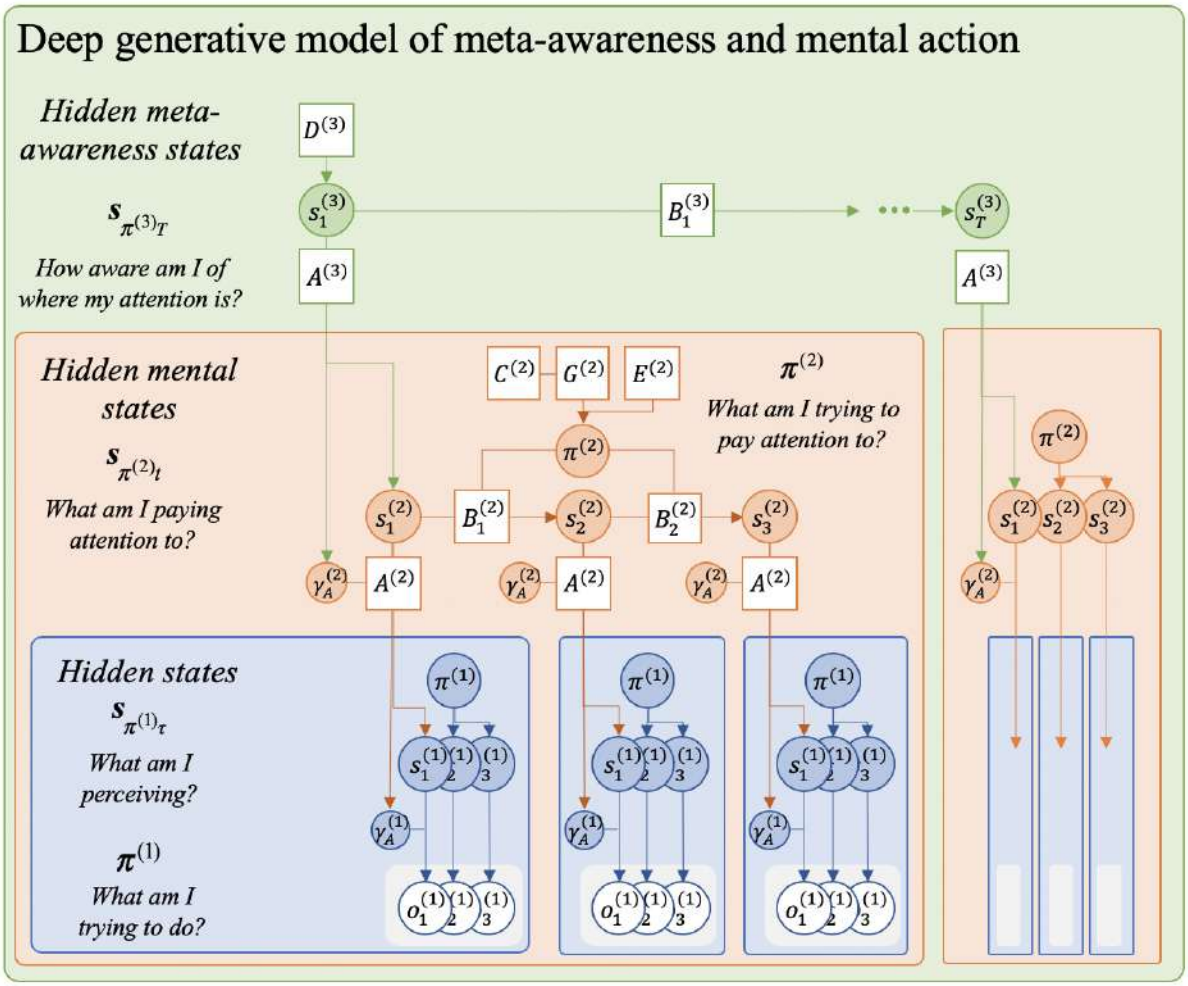

Figure 5: A probabilistic graphical model showing a deep generative model with three hierarchical levels of state inference. Higher-level states condition the likelihood mapping precision at the level below. Attentional states, $s^{(2)}$, modulate the confidence in sensory observations and meta-awareness states, $s^{(3)}$, modulate the confidence in metacognitive observations.

To summarise, this paper makes three novel contributions to parametrically deep active inference models of metacognition. The first contribution is to stipulatively define attentional and meta-awareness *states*, as opposed to *processes* per se. The second is to provide a formal definition of (i.e., implicit) attentional *control* via an account of mental policy selection. The third is to formally demonstrate the relationship between meta-awareness states and the capacity for deliberate (i.e., explicit) attentional control. The resulting model is presented in Figure 5. In what follows, we illustrate belief updating using the above inference architectures to illustrate the emergence of mindful and metacognitive phenomena.

3. Results

This section provides numerical (simulation) results to show how the model described in the previous section engenders and opacity-transparency phenomenology. This provides a formal account of meta-awareness, and also a model for attentional control – formalized as state-dependent, precision control. Our strategy will be to simulate the three hierarchical levels of inference of the model, thereby providing a proof of concept that focused attention or distraction can be reproduced by belief updating in this model.

3.1. Level 1: Attention, opacity, and awareness

With the above architecture in place, we can begin to examine how the process of sustained attentional control unfolds. Beginning at the first hierarchical level, we can simulate the effect of equipping a generative model with attentional states by examining numerically how varying precision on $\mathbf{A}^{(1)}$ impacts the dynamics of perceptual posterior state estimation.

In the simulation below, depicted in Figure 6, two different active inference agents are shown in a passive oddball paradigm. In such a paradigm, the agent is presented with a repeating standard stimulus, followed by the occasional deviant stimulus (the oddball). This is meant to reproduce the changes in the content of perception during focused attention and distraction. Sensory habituation is not modelled here. At each timestep, the agent infers the latent cause of their observations, i.e., the actual stimuli that was presented. The first agent is endowed with a high level of sensory precision (high precision $\boldsymbol{\gamma}_{\mathbf{A}}$ over the likelihood mapping) and provides a simple illustration of a participant who is paying attention to what they are perceiving. The second agent, on the other hand, has a lower precision over $\mathbf{A}^{(1)}$ and, for our purposes represents a distracted participant. Mathematically speaking, changing the expected precision $\boldsymbol{\gamma}_{\mathbf{A}}$ can have the same effect as updates in individual elements of $\mathbf{A}^{(1)}$ in response to incoming evidence, which can render the likelihood mapping more or less informative. The core difference is that the expected precision applies to all the elements of the likelihood mapping; thereby providing empirical constraints on the mapping – and a different kind of statistical structure to the generative model. If this structure is apt to describe the high-order statistics of the sensorium at hand, it will promote self-evidencing via a minimisation of complexity – a property we exploit in the current paper.

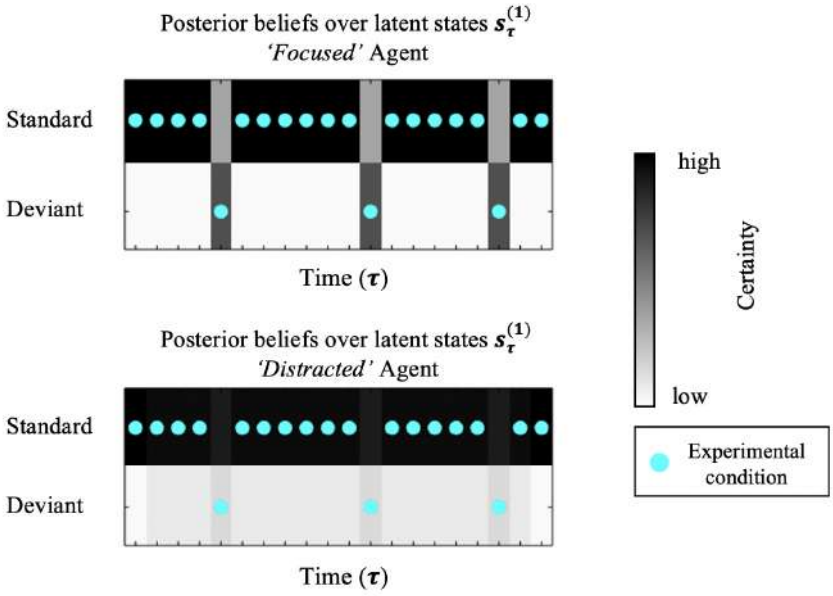

Figure 6. Simulation of two active inference agents attending an oddball paradigm. Each agent is equipped with different precisions over the likelihood mapping $\mathbf{A}^{(1)}$. The top part of the figure depicts the performance of an agent with precise likelihood beliefs on the passive oddball paradigm; and the bottom part depicts the performance of an agent with low-precision likelihood beliefs. Greyscale shading represents the certainty of posterior beliefs. The high-precision agent is generally more confident that its observations map onto states.

The focused agent is able to update their beliefs about what they are seeing immediately upon presentation of the deviant stimulus. The distracted agent does not have enough confidence in their observations to update their beliefs as quickly, and as a result, their beliefs about hidden states are not adjusted when the deviant is presented. This illustrates a ubiquitous aspect of precision control and attentional gain; namely, precision plays the role of a *rate constant* in evidence accumulation and consequent belief updating. In other words, when attending to a particular stream of information, you will converge on posterior beliefs more quickly because certain aspects of sensory information are afforded more precision and have a greater influence on belief updating, at higher levels of hierarchical inference.

3.2. Level 2: Simulating inference of hidden attentional states: the cycle of focused attention and distraction

Given the hierarchical model described above, precisions at this first level can be modulated by higher-level attentional states (i.e., $s^{(2)}$). We now introduce a modified version of the oddball paradigm described above, which implements a form of mental action. In this modified oddball paradigm, the agent's task is to move their attentional focus to the stimuli (the repeating uniform stimulus) and to maintain focus on this object, thereby remaining in a specific attentional state (*Focused*).

Extant work on the phenomenological dynamics of focused attention tasks has shown that, in general, an individual will cycle back and forth between two states: remaining focused and becoming distracted (Hasenkamp et al., 2012; Lutz et al., 2008). This cycle goes through four

distinct phases. To begin with, the individual is focused on a particular task or stimulus and successfully maintains their attention on a focal point or object. We label this attentional state '*Focused*' or '*F*'. At some point, they will inevitably become distracted, transitioning to an attentional state that we label as '*Distracted*' or '*D*'. Crucially, this state transition is, at least initially, unknown to the individual for a short period of time. This period is known as *mind wandering*, a state of being distracted whilst also unaware of their being distracted. Eventually, the individual realises they are no longer focused and become aware of having become distracted, a moment we label '*Aware of distraction*' or '*A*', which then prompts them to redirect their focus to the task at hand (i.e., returning to attentional state *F*). For simplicity, the final phase of this cycle, cognitive reappraisal of the distractor (e.g., 'just a thought'), is not modelled here.

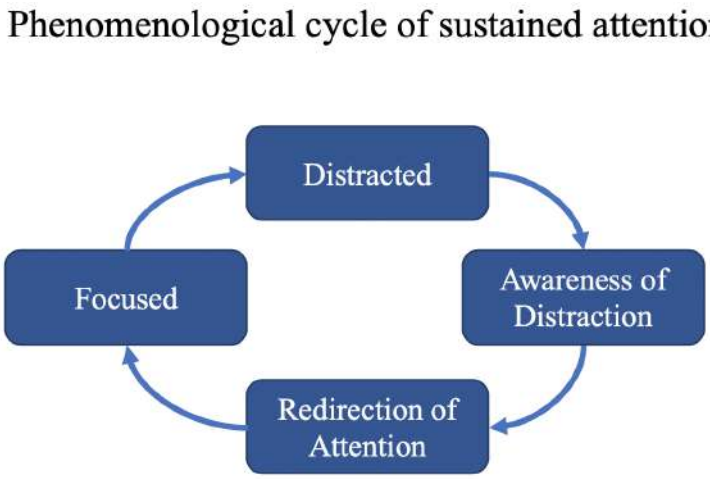

Figure 7. Diagram of the phenomenological cycle that occurs during sustained attention tasks. The process cycles through being focused, becoming distracted, becoming aware of the distraction, and then refocusing.

Asking a participant to remain focused on a particular stimulus is equivalent, in this scheme, to asking them to maintain a higher-order attentional state (*Focused*). In order to remain focused and notice whether they have become distracted, the agent must continuously infer which attentional state they are in. By defining attentional states as controllable higher-order latent states (i.e., as states that can be selected through policy selection at the second level), we effectively allow the agent to control what she is attending to.

We can test whether this architecture gives rise to the attentional dynamics of attention and distraction we might expect. This well documented phenomenological cycle of focused attention, described above, emerges organically from simulations, when we cast attentional states as a higher-level latent states that the agent must infer.

In the simulation reported below, an active inference agent infers which attentional state it is currently in. Two methodological points need to be mentioned: First, we have built in a preference for the agent to observe itself in the *Focused* state (this is built into the prior preference or $\mathbf{C}^{(2)}$ matrix). Thus, the agent expects itself to be in the *Focused* state and will

engage in active inference such that this expectation or preference is fulfilled: if the agent infers that it has become *Distracted*, it will enact a policy to return it to the *Focused* state. Second, we have programmed a *generative process* that regulates the 'actual' state transitions that the agent undergoes. Under this generative process, there is some prespecified probability that the agent will transition from one attentional state to the other spontaneously, e.g., from *Focused* to *Distracted*. This is implemented quite simply as a forced distraction: at timestep 10, the true state shifts from *F* to *D*.

The results are shown in Figure 8. Here, at timestep 10, the agent is presented with a distractor, causing the latent attentional state to transition to *Distracted*. Since the agent must *infer* their attentional state, however, this fact is not immediately inferred. Metacognitive observations of the agent's attentional state results in a prediction error that causes the agent to revise their beliefs after a few observations, with increasing confidence.

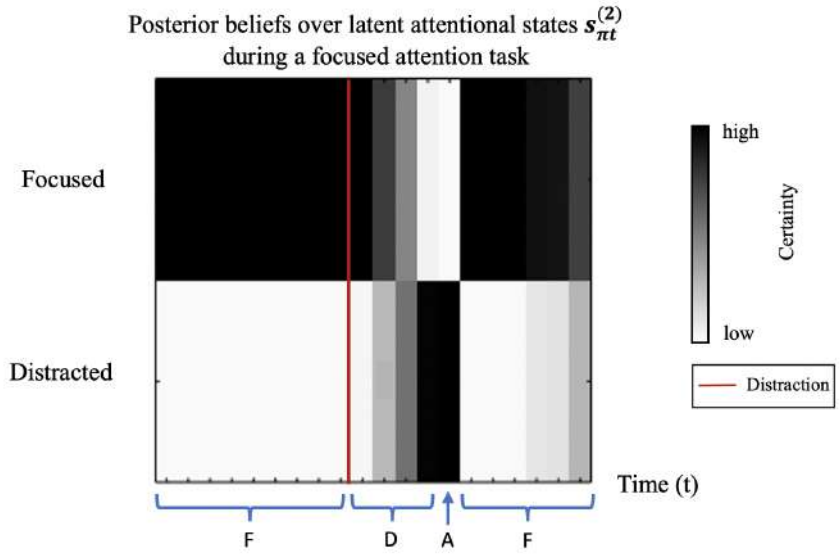

Figure 8: Simulation of an agent trying to remain in a focused attentional state (F). Numerical demonstration of the cycle of distraction (D), awareness of distraction (A) and redirecting focus. The active inference agent is inferring their own attentional state at each time step, which can be either *Focused* or *Distracted.* At the beginning of the task the agent is *Focused.* At t = 10, the agent is forcefully distracted. After a few moments, the agent realises they have become distracted and, given that they would prefer to observe themself in the *Focused* state, selects a mental policy, $\boldsymbol{\pi}^{(2)}$, to change their attentional state.

Here, the focused stage (*F*) is the period during which the agent's beliefs align with the true (*Focused*) attentional state. The mind wandering stage is well captured by the numerical results, as the moments when the true attentional state has transitioned to *Distracted*, whilst the agent's beliefs have not yet been updated (the agent still believes it is *Focused*). The agent is effectively unaware of their distracted state: its 'mind' has wandered. Over the following time steps, the agent collects enough evidence (i.e., metacognitive observations of their attentional state) to update their beliefs and to 'realise' that their attentional state has shifted (to *D*). Finally, now aware of their being distracted, the agent performs a mental action at the second level, to transition the attentional state back to *Focused*.

It is important for the reader to keep in mind that the *Distracted* versus *Focused* states do not require any objective existence independent of the observer. The only thing that matters is that the observer believes these states exist and gathers evidence for or against them. In this sense, the 'true' state is defined by what the perceiver counts as evidence for that particular state. Therefore, we do not indicate any 'true' states in the figures beyond the lowest level of the hierarchy, as these are the only states under experimental control – the rest being controlled by the perceiving mind. As a result of Bayesian model-averaging, the perceiving mind will always live somewhere on the spectrum between the two extremal states of being entirely focused and entirely distracted.

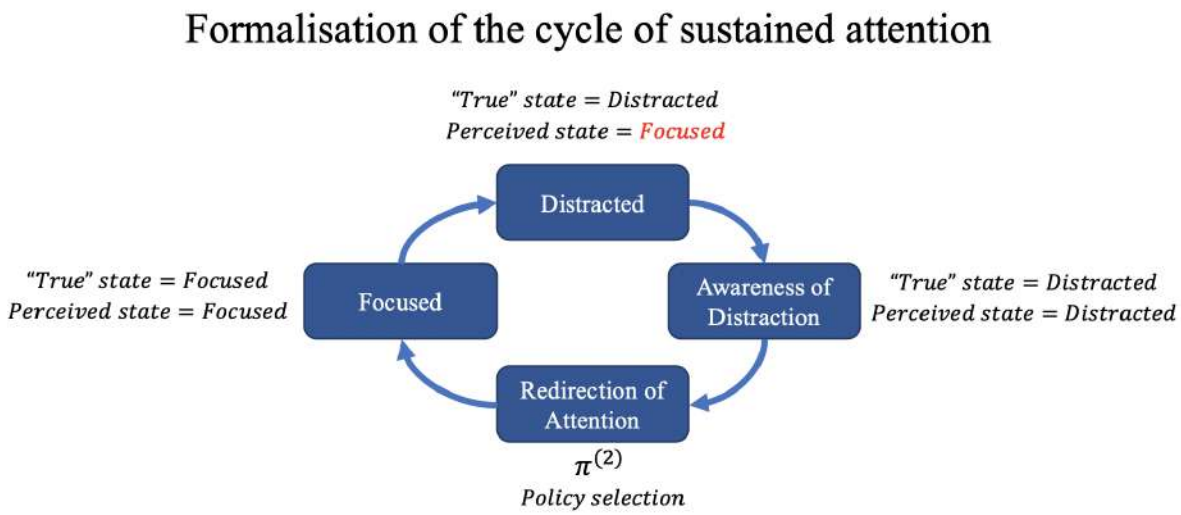

Figure 9: Illustration of the computational conditions associated with each phenomenological stage of sustained attention. This figure depicts schematically the main phases of the cycle of focused attention and distraction. Here, being distracted (i.e., mind wandering) is characterised as the period following the shift in the true latent attentional state from *Focused* to *Distracted*, but before the agent has updated their beliefs to align with the true latent attentional state; i.e., the period in which the agent believes that it is *Focused* while it is actually *Distracted*.

3.3. Level 3: Simulating the impact of changes in meta-awareness states on attentional control

As discussed above, the ability to maintain attentional focus and to quickly become aware of distractions has been associated with the level of meta-awareness of the individual (Mrazek et al., 2013). We now examine the impact of changing meta-awareness states on attentional state inference dynamics. We demonstrate that meta-awareness-driven state-dependent control arises naturally from casting meta-awareness states as third-order states that modulate the likelihood precision of second-order attentional states; much like attentional states modulate the precision of first-order perceptual states.

In Figure 10, two simulated agents perform the same focused attention task as just described, with one notable change: we introduce a second distractor at timestep 20, in addition to the

one at timestep 10. In this setup, one agent is endowed with a low confidence in the metacognitive observations of its own attentional state, i.e. an imprecise likelihood mapping $\mathbf{A}^{(2)}$. This case represents a situation where the agent has a low level of meta-awareness (i.e., attentional processes have low opacity). The other agent has high precision and represents an agent with a high level of meta-awareness.

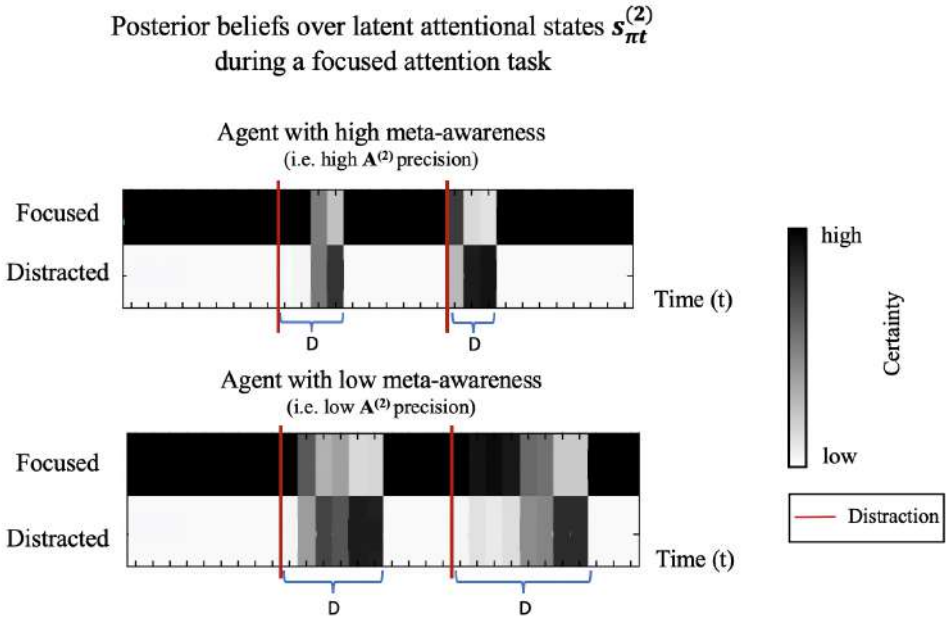

Figure 10. Simulation of two agents with differing levels of meta-awareness during a sustained attention task. This figure depicts two active inference agents, one with high precision on $\mathbf{A}^{(2)}$, the other with low precision on $\mathbf{A}^{(2)}$, that are engaged in a modified version of the oddball stimulus described in the text. In this experiment, both agents are forcefully distracted at t = 10 and t = 20. Note that the period of state evidence accumulation is increased (i.e., longer mind wandering) for the agent with low meta-awareness.

We report that the decreased precision of $\mathbf{A}^{(2)}$ (i.e., reduced opacity of the attentional states) results in an extended period of mind wandering before the agent accumulates enough evidence to realise they have become distracted. The reduction of the duration of mind wandering, due to stronger meta-awareness modulation and increased precision afforded to attentional states, is an established relationship in attentional phenomenology, particularly in relation to focused attention meditation practices (Mrazek et al., 2013). This relationship is illustrated here as an organic dynamic that emerges naturally from the hierarchical architecture of higher-level states encoding precisions at lower levels. This completes our numerical analysis.

4. Discussion and directions for future research

4.1. Expanding the model to other parameters

The model presented makes it possible to simulate, under a single framework, both physical (overt) actions and mental (covert) actions – in effect providing a single model of perceptual inference and behaviour that can provide a computational bridge between mental and

embodied life. This work provides a principled approach to modelling complex behaviours in tasks that require both motor action selection (e.g., saccadic movements) and the deliberate deployment of attentional resources (e.g., paying attention to a particular stimulus to the exclusion of another). Our model also provides insights into the form of cognitive architectures which may be required to support the emergence of metacognitive monitoring (i.e., phenomenological opacity of mental states) and control.

We have focused on attentional processes that are implicate the likelihood mapping or **A** matrix. The natural next step is to consider the implications of generalising this treatment, to model the access to – and control of – the precision of other parts of the generative model; e.g., the precision of beliefs about state transitions (**B**), preferences over outcomes (**C**), prior beliefs over states (**D**), prior beliefs about policies (**E**), and the expected variational free energy itself (**G**; as in Hesp et al., 2020). Mathematically, it is perfectly valid to condition the prior precisions associated with any part of the generative model on a higher-level state. Figure 11 depicts the structure of a generative model that might enable this.

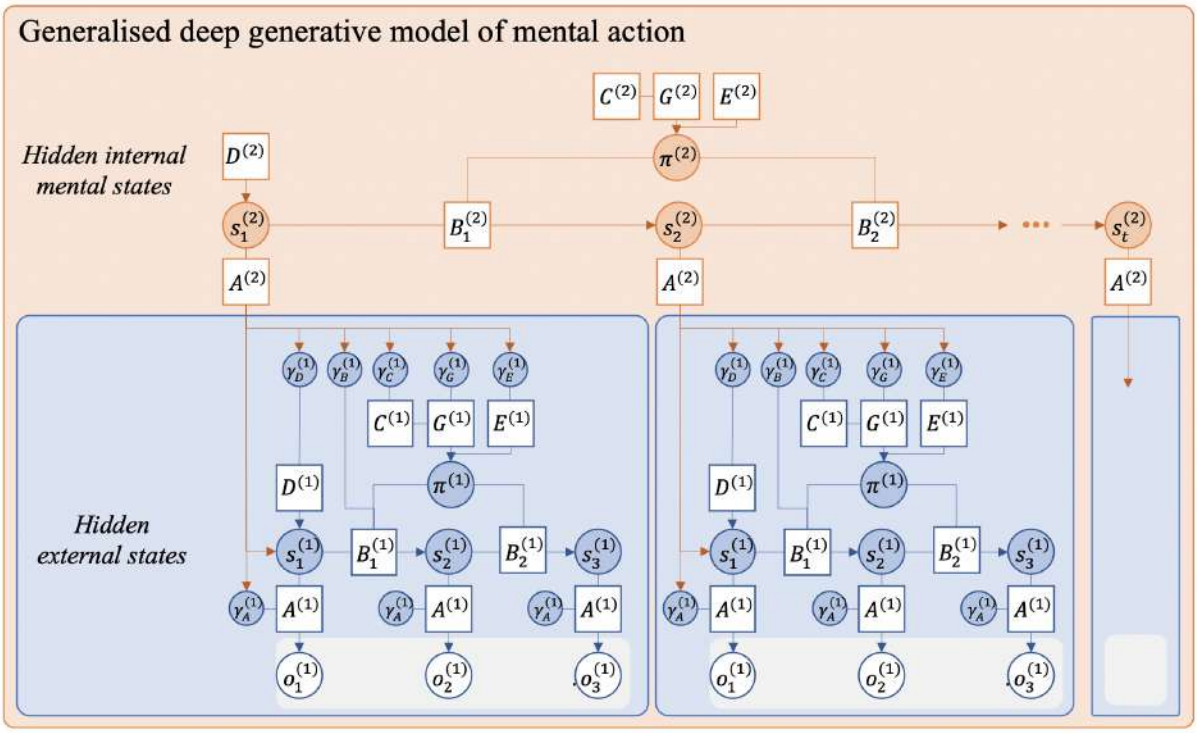

Figure 11. A Bayes graph of a deep generative model of mental action generalised over all precision parameters. With this architecture, higher-level states $s^{(2)}$ contain multiple factors modulating all lower-level precisions. This structure provides a direction towards the formalisation of a wide phenomenology of mental states and policies ($\boldsymbol{\pi}^{(2)}$). Adapted from a template of Figure 4 from Hesp et al. (2020).

The result of this innovation is a diversification of the mental actions available to the agent. Thus, higher-level states $s^{(2)}$ stand for hidden mental states generally and are not restricted to modelling attentional states. Mapping the precision of the other parts of the generative model onto dimensions of phenomenological experience is beyond the scope of this paper. We anticipate that this strategy could provide a general framework to conceptualise different styles of metacognitive regulations and mental actions. For instance, this architecture may

prove to be sufficient to model and compare some computational features of a dual versus non-dual forms of meta-awareness in mindfulness meditation (Dunne et al. 2019). It could also provide a computational scaffold for examining how different cultural and social settings shape metacognitive styles in different ways (Proust & Fortier, 2018), with implications for our understanding of metacognitive development and cross-cultural variability in psychiatric symptomatology. An exciting avenue will be to use this approach to refine and extend the modelling of various psychiatric conditions such as mood disorders, or metacognitive disorders (Kiverstein et al., 2020).

This modelling strategy, we emphasise, is completely general – and is not restricted to modelling attentional processes. It shares a commitment to understanding perception in terms of inference (Fleming, 2020; Gregory, 1968, 1980; Helmholtz, 1866/1962); and, in particular, the high order aspects of perceptual inference, such as that celebrated in hierarchical treatments of metacognition (Fleming, 2020). In terms of enactive perception (Wurtz, McAlonan, Cavanaugh, & Berman, 2011), active inference provides a principled description of how actions are deployed to reduce variational free energy; in this generic scheme, there is no fundamental difference between a simple reflex, a series of complex movements, and a mental action: c.f., the premotor theory of attention (Rizzolatti, Riggio, Dascola, & Umilta, 1987). We model mental action as a generic hidden process that is at the root of many aspects of human mental life. Other such aspects, such as emotional self-awareness and control, would be implemented in the same way (i.e., through the deployment of precision). For example, in the work of Hesp and colleagues (2020), expected precision of the action model itself (**G**) has been associated with valenced responses. Our model is generic and adaptable; but it is not currently implemented to cover other kinds of mental action. We intend to explore such directions in future work.

4.2. Implications for computational psychiatry

In active inference, where action selection becomes part of the inference process, precisions are ubiquitous and multifaceted. Namely, sensory precisions control the likelihood mapping (between hidden states and sensory consequences), while other precisions control the transitions among hidden states, and others still control beliefs about policies. These three kinds of precision have been associated with cholinergic, noradrenergic, and dopaminergic neurotransmission, respectively (Sterzer et al., 2018). Such neurotransmitters have been hypothesised to function as reliable signals indicating various types of confidence associated with inferential processes. Hence, from both a psychological and neurobiological perspective, the way precisions themselves are tuned appears as a central question to explain various psychiatric conditions.

In particular, aberrant precision tuning has been pointed to be very relevant for understanding hallucinations (Powers et al., 2017) and psychosis (Sterzer et al., 2018). In recent years, more and more studies have provided formal, computational explanations for an increasing number of psychiatric symptoms such as those of schizophrenia (Sterzer et al., 2019; Wacongne, 2016; Benrimoh et al., 2018), autism (Palmer et al., 2017; Robic et al., 2015; Sapey-

Triomphe et al., 2018; Constant et al., 2018), and depression (Stephan et al., 2016). Such precisions can be linked to hyperparameters of generative models that explain the behaviours of subjects. Given the relationship between these precision terms and various neurotransmitters, we could directly simulate the effects of pharmacological treatment on the neurocognitive and behavioural dynamics of a particular patient.

The hierarchical model we introduced in this paper explicitly tackles the processes by which the various precision parameters are inferred. In line with recent work by Hesp and colleagues (2020) and in contrast with previous deep models, here precision parameters are endowed with hierarchical priors. Their deployment is what we refer to as a covert or mental action. Such a formulation opens up broad perspectives for capturing the essence of a psychiatric condition. It suggests that through behavioural or neurophysiological observations, one might be able to characterise an individual profile along relevant psychological dimensions. In other words, the proposed model architecture is among the first to make explicit the conscious awareness and control of ubiquitous and essential model precisions. Our work thus helps pave the way for computational phenotyping and precision psychiatry (Friston, 2017).

The distinction between – on the one hand – accessing or perceiving one's own mental states and – on the other hand – controlling them is crucial here. It is the control aspect that is novel in our approach to metacognition in the computational modelling tradition. Indeed, computationally speaking, one could argue that hierarchical generative models (e.g., empirical Bayes) have been around for a long time, and that they are useful in allowing us to implement perception and learning as inference in a way that enables to infer the (deep) causal structure of the environment. This part is not exactly novel. Such a hierarchical structure underlies mainstream approaches in machine learning via deep neural networks (e.g., convolutional neural network and recurrent neural networks, which build up a hierarchical model with hidden layers pertaining to different features at different scales). What is novel about our model is the coupling of this deep inferential architecture with *action and its control*; that is, we do not speak merely of deep inference but of deep *active* inference. Our modelling strategy offers us a means of controlling previously uncontrolled parameters through the top-down deployment of precisions.

4.3. Towards a formal or computational neurophenomenology

Another future possible application of the present framework is to provide a formal tool to explore, theoretically and experimentally, the *naturalisation of phenomenology* as proposed by the proponents of neurophenomenology (Lutz, 2002; Lutz & Thompson, 2003; Petitot, 1999; Roy et al., 1999; Varela, 1996, 1997). Naturalising phenomenology is a scientific research program that aims to characterise the mind-brain system on its own terms, as it were; the way that it appears to itself for itself, as a subject of experience, rather than only focusing on it as a mere thing (Roy et al., 1999). What is at stake here is how best to characterise the relation between first-person data obtained from phenomenological accounts of lived experience to third-person cognitive and neuroscientific accounts. For instance, how might

one relate the direct lived experience of watching a beautiful sunset to its physiological manifestation?

Because the proponents of neurophenomenology believe that first-person experience opens onto a field of phenomena that is irreducible to any other, they typically acknowledge the epistemological importance of bridging the kind of knowledge gleaned from first-person (phenomenological) and third-person (ordinary) data (Petitot, 1999). Our take on neurophenomenology follows the route of formal neurophenomenology (Lutz, 2002; Petitot, 1999; Varela, 1997), which aims to bridge the gap between these kinds of data via formal and metaphysically neutral levels of description provided by mathematics and computational work. Formal neurophenomenology formalises the aspects of lived experience that are made accessible by phenomenological description and models of the inferential processes that enable the emergence of target phenomenologies. The strategy of neurophenomenology (Varela, 1996) then is to build and validate an integrative model of conscious experience based on "mutual constraints" between the domain of phenomena revealed by experience, the domain of neurophysiological states that are measured by cognitive neurosciences, and the domain circumscribed by formal models (Varela, 1996, 1997).

The pioneers of neurophenomenology had sought to bridge this gap using formal models and analytical tools from dynamical systems theory. From our point of view, these can be seen as anticipating, without the same degree of formal precision, several core principles of the active inference framework. The formalism available at the time was not yet well enough equipped to model explicitly those subtle phenomenological constructs such as transparency, opacity, meta-awareness, mental action. In this view, the present model could extend and complement this earlier endeavour.

From the point of view of formal neurophenomenology, deep active inference models can act as maps of the processes and factors at play in the emergence and dynamics of lived or conscious experience. Indeed, the model presented here first started from the phenomenology of focused attention, and emerged as a model of the architecture of beliefs and inferences that would make such a phenomenology possible, and interpretable as such an experience. The degree of complexity of the model was commanded and constrained by the target phenomenology; in this case, we needed three hierarchical levels. The general form of this kind investigation – using active inference directly as a means of formalising the basic requirements that can explain the structure of lived experience – may point towards an interesting path for formal neurophenomenology.

5. Conclusion

The aim of this paper was to begin moving towards a formal neurophenomenology of metacognition based on deep active inference. Understanding metacognition is critical to the study of human beings, since it is perhaps the most characteristic facet of the human experience. We proposed a formal model and computational proof of principle that

operationalises some central aspects of metacognitive monitoring and control. We used the modelling and mathematical tools of the active inference framework to construct an inferential architecture (a generative model) for meta-awareness of, and metacognitive control of, attentional states. This model consists of three nested levels, which afforded respectively (1) perception of the external environment, (2) perception of internal attentional states (focused versus distracted), and (3) perception of meta-awareness states (aware versus unaware). This architecture enables the modelling of higher-level, mental (covert) action, granting the agent some control of its own attentional processes. We replicated in silico some of the more crucial features of meta-awareness, including its phenomenology and relationship to metacognitive control.

6. References

Allen, M., Levy, A., Parr, T., & Friston, K. J. (2019). In the Body's Eye: The Computational Anatomy of Interoceptive Inference. *BioRxiv*. https://www.biorxiv.org/content/10.1101/603928v1.abstract

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

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

Bernstein, A., Hadash, Y., Lichtash, Y., Tanay, G., Shepherd, K., & Fresco, D. M. (2015). Decentering and Related Constructs: A Critical Review and Metacognitive Processes Model. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *10*(5), 599–617.

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14*(4), 411–427.

Constant, A., Bervoets, J., Hens, K., & Van de Cruys, S. (2018). Precise Worlds for Certain Minds: An Ecological Perspective on the Relational Self in Autism. *Topoi. An International Review of Philosophy*. https://doi.org/10.1007/s11245-018-9546-4

Dahl, C. J., Lutz, A., & Davidson, R. J. (2015). Reconstructing and Deconstructing the Self in

Three Families of Meditation. *Trends in Cognitive Sciences*.

Eberth, J., & Sedlmeier, P. (2012). The Effects of Mindfulness Meditation: A Meta-Analysis. In *Mindfulness*. https://doi.org/10.1007/s12671-012-0101-x

Farb, N., Daubenmier, J., Price, C. J., Gard, T., Kerr, C., Dunn, B. D., Klein, A. C., Paulus, M. P., & Mehling, W. E. (2015). Interoception, contemplative practice, and health. *Frontiers in Psychology*, *6*, 763.

Feynman, R. P. (1972). *Statistical Mechanics*. Benjamin, Reading MA, USA.

Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, 2020(1). doi:10.1093/nc/niz020

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1280–1286.

Fletcher, L., & Hayes, S. C. (2005). Relational frame theory, acceptance and commitment therapy, and a functional analytic definition of mindfulness. *Journal of Rational-Emotive and Cognitive-Behavior Therapy: RET*, *23*(4), 315–336.

Fox, K. C. R., Dixon, M. L., Nijeboer, S., Girn, M., Floman, J. L., Lifshitz, M., Ellamil, M., Sedlmeier, P., & Christoff, K. (2016). Functional neuroanatomy of meditation: A review and meta-analysis of 78 functional neuroimaging investigations. In *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2016.03.021

Friston, K. J. (2011). *Embodied inference: or" I think therefore I am, if I am what I think."* https://psycnet.apa.org/record/2014-14659-005

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society, Interface / the Royal Society*. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J. (2017). Precision Psychiatry [Review of *Precision Psychiatry*]. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *2*(8), 640–643.

discovery.ucl.ac.uk.

Friston, K. J. (2019). A free energy principle for a particular physics. In *arXiv [q-bio.NC]*. arXiv. http://arxiv.org/abs/1906.10184

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

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference: a process theory. *Neural Computation*, *29*, 1–49.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210.

Friston, K. J., & Parr, T. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society, Interface / the Royal Society*. https://doi.org/10.1016/j.neuron.2005.04.026

Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational Nosology and Precision Psychiatry. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 2–23.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society B*, 171, 179-196.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B*, 290, 181-197.

Grush, R. (2005). Internal models and the construction of time: generalizing from state estimation to trajectory estimation to address temporal features of perception, including temporal illusions. *Journal of Neural Engineering*, *2*(3), S209–S218.

Hasenkamp, W., Wilson-Mendenhall, C. D., Duncan, E., & Barsalou, L. W. (2012). Mind wandering and attention during focused meditation: A fine-grained temporal analysis of fluctuating cognitive states. *NeuroImage*.

https://doi.org/10.1016/j.neuroimage.2011.07.008

Helmholtz, H. (1866/1962). Concerning the perceptions in general (J. Southall, Trans.). In *Treatise on Physiological Optics* (Vol. III). New York: Dover.

Hesp, C., Smith, R., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2020). Deeply Felt Affect: The Emergence of Valence in Deep Active Inference. *PsyArXiv*. https://doi.org/10.31234/osf.io/62pfd

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259-285. doi:10.1111/nous.12062

Husserl, E. (1927). Phenomenology. *Encyclopaedia Britannica*, *14*, 699–702.

Jamieson, G. A. (2016). A unified theory of hypnosis and meditation states: The interoceptive predictive coding approach. . In A. Raz & M. Lifshitz (Eds.), *Hypnosis and Meditation: Towards an Integrative Science of Conscious Planes* (p. 313–342). Oxford University Press. Kiverstein, J., Miller, M., & Rietveld, E. (2020). How Mood Tunes Prediction: A Neurophenomenological Account of Mood and its Disturbance in Major Depression. *Neuroscience of Consciousness*.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Lebois, L. A. M., Papies, E. K., Gopinath, K., Cabanban, R., Quigley, K. S., Krishnamurthy, V., Barrett, L. F., & Barsalou, L. W. (2015). A shift in perspective: Decentering through mindful attention to imagined stressful events. *Neuropsychologia*, *75*, 505–524.

Lecaignard, F., Bertrand, O., Caclin, A., & Mattout, J. (2020). Adaptive cortical processing of unattended sounds: neurocomputational underpinnings revealed by simultaneous EEG-MEG. In *bioRxiv* (p. 501221). https://doi.org/10.1101/501221

Limanowski, J., & Friston, K. J. (2018). “Seeing the Dark”: Grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.00643

Limanowski, J., & Friston, K. J. (2020). Attenuating oneself: An active inference perspective on “selfless” experiences. *Philosophy and the Mind Sciences*, *1*(6).

Lutz, A. (2002). Toward a neurophenomenology as an account of generative passages: a first empirical case study. *Phenomenology and the Cognitive Sciences*, *1*(2), 133–167.

Lutz, A., Jha, A. P., Dunne, J. D., & Saron, C. D. (2015). Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *The American Psychologist*. https://doi.org/10.1037/a0039585

Lutz, A., Mattout, J., & Pagnoni, G. (2019). The epistemic and pragmatic value of non-action: a predictive coding perspective on meditation. In *Current Opinion in Psychology*. https://doi.org/10.1016/j.copsyc.2018.12.019

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. In *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2008.01.005

Lutz, A., & Thompson, E. (2003). Neurophenomenology Integrating Subjective Experience and Brain Dynamics in the Neuroscience of Consciousness. *Journal of Consciousness Studies*, *10*(9-10), 31–52.

Lysaker, P. H., Keane, J. E., Culleton, S. P., & Lundin, N. B. (2020). Schizophrenia, recovery and the self: An introduction to the special issue on metacognition. *Schizophrenia Research. Cognition*, *19*, 100167.

Manjaly, Z.-M., & Iglesias, S. (2019). A Computational Theory of Mindfulness Based Cognitive Therapy from the “ Bayesian Brain ” P erspective. *PsyArXiv*.

Merleau-Ponty, M. (1945). Phénoménologie de la perception. *Paris 1945*. http://visions419.rssing.com/chan-24754465/all_p9.html

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*(4), 353–393.

Metzinger, T. (2004). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B., & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering. *Psychological Science*, *24*(5), 776–781.

Nelson, T. O. (1996). Consciousness and metacognition. *The American Psychologist*, *51*(2), 102–116.

Pagnoni, G. (2019). The contemplative exercise through the lenses of predictive processing: A promising approach. *Progress in Brain Research*, *244*, 299–322.

Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin*, *143*(5), 521–542.

Papies, E. K., Barsalou, L. W., & Custers, R. (2012). Mindful Attention Prevents Mindless Impulses. *Social Psychological and Personality Science*, *3*(3), 291–299.

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

Petitot, J. (1999). *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*. Stanford University Press.

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600.

Proust, J., & Fortier, M. (2018). *Metacognitive Diversity: An Interdisciplinary Approach*. Oxford University Press.

Ramstead, M. J. D. (2015). Naturalizing what? Varieties of naturalism and transcendental phenomenology. *Phenomenology and the Cognitive Sciences*, *14*(4), 929–971.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16.

Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205.

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: active inference is enactive inference. *Adaptive Behavior*, 1059712319862774.

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

Reder, L. M., & Schunn, C. D. (2014). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In *Implicit memory and metacognition* (pp. 57–90). Psychology Press.

Rizzolatti, G., Riggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A), 31-40. doi:10.1016/0028-3932(87)90041-8.

Robic, S., Sonié, S., Fonlupt, P., Henaff, M.-A., Touil, N., Coricelli, G., Mattout, J., & Schmitz, C. (2015). Decision-making in a changing world: a study in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *45*(6), 1603–1613.

Roy, J.-M., Petitot, J., Pachoud, B., & Varela, F. J. (1999). Beyond the gap: An introduction to naturalizing phenomenology. In *Naturalizing phenomenology* (pp. 1–83). Stanford University Press.

Sapey-Triomphe, L.-A., Sonié, S., Hénaff, M.-A., Mattout, J., & Schmitz, C. (2018). Adults with Autism Tend to Undermine the Hidden Environmental Structure: Evidence from a Visual Associative Learning Task. *Journal of Autism and Developmental Disorders*, *48*(9), 3061–3074.

Schooler, J. W., & Smallwood, J. (2009). Metacognition. In *Oxford Companion to Consciousness* (pp. 438–442). Oxford University Press.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences*, *15*(7), 319–326.

Schwartenbeck, P., & Friston, K. J. (2016). Computational Phenotyping in Psychiatry: A Worked Example. *eNeuro*, *3*(4). https://doi.org/10.1523/ENEURO.0049-16.2016

Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., & Kunze, S. (2012). The psychological effects of meditation: A meta-analysis. *Psychological Bulletin*. https://doi.org/10.1037/a0028168

Segal, Z. V., & Teasdale, J. (2018). *Mindfulness-Based Cognitive Therapy for Depression, Second Edition*. Guilford Publications.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic Self-efficacy: A Metacognitive Theory of Dyshomeostasis-Induced Fatigue and Depression. *Frontiers in Human Neuroscience*, *10*, 550.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*, *84*(9), 634–643.

Sterzer, P., Voss, M., Schlagenhauf, F., & Heinz, A. (2019). Decision-making in schizophrenia: A predictive-coding perspective. *NeuroImage*, *190*, 133–143.

Tang, Y.-Y., Hölzel, B. K., & Posner, M. I. (2015). The neuroscience of mindfulness meditation. *Nature Reviews. Neuroscience*, *16*(4), 213–225.

Tellegen, A., & Atkinson, G. (1974). Openness to absorbing and self-altering experiences ("absorption"), a trait related to hypnotic susceptibility. *Journal of Abnormal Psychology*, *83*(3), 268–277.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., &

Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological Review*, *121*(4), 649–675.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*.

Varela, F. J. (1997). *The Naturalization of Phenomenology as the Transcendence of Nature: Searching for generative mutual constraints*.

Wacongne, C. (2016). A predictive coding account of MMN reduction in schizophrenia. *Biological Psychology*, *116*, 68–74.

Wetherell, J. L., Hershey, T., Hickman, S., Tate, S. R., Dixon, D., Bower, E. S., & Lenze, E. J. (2017). Mindfulness-Based Stress Reduction for Older Adults With Stress Disorders and Neurocognitive Difficulties. In *The Journal of Clinical Psychiatry* (Vol. 78, Issue 7, pp. e734–e743). https://doi.org/10.4088/jcp.16m10947

Wiese, W. (2017). Predictive processing and the phenomenology of time consciousness. Philosophy and Predictive Processing. Frankfurt Am Main: MIND Group. Https://d-nb.info/1Predictive processing and the phenomenology of time consciousness. Philosophy and Predictive Processing. Frankfurt Am Main: MIND Group. P

Wiese, W. (2017). Predictive processing and the phenomenology of time consciousness. *Philosophy and Predictive Processing. Frankfurt Am Main: MIND Group*. https://d-nb.info/1135300135/34

Wurtz, R. H., McAlonan, K., Cavanaugh, J., & Berman, R. A. (2011). Thalamic pathways for active vision. *Trends in Cognitive Sciences*, 5(4), 177-184.