The computational unconscious: Adaptive narrative control, psychopathology, and subjective well-being

Authors
George Deane 0,1*
Jonas Mago 2, 3, 4,
Katerina Fotopolou 5
Matthew D. Sacchet 6
Robin Carhart-Harris 7,8
Lars Sandved-Smith 9

Affiliations
0. Department of Philosophy, University of Montreal
1. MILA - Quebec AI Institute
2. Integrated Program in Neuroscience, McGill University
3. Division of Social and Transcultural Psychiatry, Department of Psychiatry, McGill University, Montreal, Canada.
4. Wellcome Centre for Human Neuroimaging, University College London, London, UK.
5. Research Department of Clinical, Educational and Health Psychology, University College London, UK.
6. Meditation Research Program, Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States
7. Departments of Neurology and Psychiatry, University of California San
Francisco
8. Centre for Psychedelic Research, Division of Brain Sciences, Imperial College London
9. Monash Centre for Consciousness and Contemplative Studies, Monash University, 29 Ancora Imparo Way, Clayton, VIC, 3800, Australia

***Corresponding author:** georgejwdeane@gmail.com

Abstract

This paper introduces the notion of adaptive narrative control, a conception of how subpersonal computational processes shape the contents of conscious experience to realize adaptive behavior. We unpack the implications of the theory for understanding the computational mechanisms underwriting psychopathology and improvements in subjective well-being associated with psychedelic therapy and meditation. The core idea of adaptive narrative control is that systems equipped with an 'attention schema' — a model of its own attentional states and how attentional states can be controlled — can come to anticipate not only the epistemic implications of certain attentional states, but also the pragmatic consequences, such as how certain attentional states potentiate certain affective responses. In anticipating affective states, the system is able to regulate affective states through 'mental action' — the endogenous control of attention. We argue that using mental action to bias the sampling of evidence to control the 'narrative' — the upshot of inference understood to correspond to the contents of conscious experience — allows the system to regulate affective and physiological states in ways that potentiate adaptive behavior. However, it is this adaptive capacity which gives rise to the computational mechanism — 'avoidant mental action', or equivalently 'motivated inattention' — which we argue is a core mechanism underlying psychopathology. We unpack this approach within the active inference framework to provide specification of the candidate mechanisms. We argue this conception can be used to account for the rigid belief formation characterizing 'canalization' and show how the decrements in subjective well-being come as a consequence of reduced recognition and categorisation of emotions (i.e., alexithymia or impaired emotional granularity). We argue that while avoidant mental action facilitates adaptive behavior, certain environmental conditions can lead it to resulting in decreased subjective well-being and psychopathology. Our account partially echoes a Freudian perspective on the function and effects of avoidant defense mechanisms like repression, and brings into view a novel computational conception of the dynamic unconscious— the 'computational unconscious'. Finally, we explore how this conceptualisation can expand and refine the 'REBUS' model of psychedelic action and therapy, and explain some of the increments in subjective well-being associated with meditation.

Introduction

In contemporary cognitive science, the brain is often considered to be an ‘inferential engine’ — using prior probabilistic knowledge to make a statistically optimal ‘best guess’ as to the causes of incoming sensations. In each moment, probabilistic expectations are knitted together to create a coherent causal model or *narrative* to explain the incoming sensory information. These expectations span a hierarchy of interlocking timescales, where expectations of slow-moving regularities constrain expectations on faster regulaties, and vice versa. The upshot is the rich and conceptually laden world we experience that is also richly detail, where sensory information is explained across multiple levels of abstraction: the holistic perceptual experience of the face of a loved one includes everything from the host of affective reactions, to conceptual information as to their age and gender, all the way down to the finest spatiotemporal scales where a detected edge is explained as caused by a single hair.

In addition to creating a rich, multimodal model of the world, the brain also controls the flow and depth of processing of the deluge of impinging information from the outside world. This selective sampling of information is attention, affording certain streams of information greater weight or influence over the construction of experiential content or narrative. These mechanisms have been sculpted over the course of our evolutionary and developmental history to engender intelligent and adaptive behavior.

This paper introduces *adaptive narrative control theory.* Adaptive narrative control theory starts from the observation that the brain does not sample information in order to construct the most accurate or veridical world (or self) model. Rather, the allocation of attention includes pragmatic considerations: how will the ‘narrative’ — the constellation of beliefs making up the contents of consciousness in a given instant — lead to adaptive behavior. This process of narrative construction involves balancing the pragmatic and epistemic implications of actions. Where pragmatic imperatives trump epistemic imperatives, the ‘accuracy’ of the narrative suffers. The narrative is constructed on the basis of what will lead the organism to the appropriate behavior in the given context.

This simple consideration, we will argue in this paper, has important implications both for understanding the core computational mechanisms driving psychopathology and for understanding the mechanisms of psychotherapeutic interventions such as psychedelic therapy, and contemplative practices like meditation for improving subjective well-being.

Specifically, we will argue that ‘avoidant mental action’ or ‘motivated inattention’ — both understood within the apparatus of adaptive narrative control theory – provide a means of understanding the etiology of psychopathology and the formation of the formation of the ‘computational unconscious’ — which can be understood to a computational analogue of the Freudian unconscious.

The structure of the paper is as follows. In section 1, we introduce the concepts of canalization as a possible common factor (‘p-factor’) underlying diverse psychopathologies. In section 2, we introduce the broad strokes of adaptive narrative control theory. In section 3 we introduce some necessary background on the active inference framework, starting with the fundamentals of the framework and then moving on to conceptions of ‘precision’, ‘mental action’, and ‘affective inference’. In section 4, we give an account of the core mechanisms of motivated inattention. In section 5, we consider how the mechanisms of motivated inattention leads to lower subjective well-being through impairing emotional granularity. In section 6, we consider how the account connects to the Freudian notion of a ‘dynamic unconscious’. In section 7, we consider implications of this conception for understanding computational mechanisms of improved subjective well-being. Finally, in section 8, we consider psychedelic therapy and meditation as amenable to explanation in terms of these mechanisms. We conclude by considering future directions for this work.

Section 1: the p-factor and canalization

Emerging evidence, across various domains of psychopathology research, points towards a common, underlying factor often referred to as the "general factor of psychopathology," or "p-factor" (Caspi et al., 2014; but see Watts et al, 2013). The hypothesized common factor cutting across disparate psychopathologies has been proposed to have utility beyond merely explaining the high rates of comorbidity among psychiatric disorders. Notably, it has been found to predict various important life outcomes such as academic achievement, social functioning, and physical health outcomes — more effectively than individual disorders (Caspi et al., 2014).

The p-factor may explain a common vulnerability factor or susceptibility to a pathophysiological mechanism underlying multiple forms of psychopathology, thereby offering a more parsimonious conceptual framework for understanding mental health disorders. The pervasive co-occurrence of psychiatric disorders, or comorbidity, has long posed challenges to traditional categorical diagnostic models (Kessler et al., 2005). This observation supports the idea of a shared vulnerability factor, as it suggests a substantial overlap across disorders, beyond what might be expected by chance. Furthermore, transdiagnostic factors such as neuroticism, maladaptive coping strategies, poor emotion regulation, and early life stress, among others, have been shown to be associated with multiple forms of psychopathology, further bolstering the argument for a common factor (Aldao et al., 2010; Caspi et al., 2014). Furthermore, neuroimaging research has pointed to shared neural substrates among various forms of psychopathology, notably involving structures within the prefrontal cortex and the limbic system (Goodkind et al., 2015) as well as the transmodal association cortex (Sydnor et al, 2021).

Carhart-Harris et al. (2022) introduce the "canalization model of psychopathology" to explain the frequently observed general psychopathology dimension ("p") underlying comorbidity and overlap between putatively distinct mental disorders. They repurpose the concept of ‘canalization’, which was originally popularized in the 1940s as a concept within biology that describes how consistent phenotypes — the observable traits of an organism — are buffered against perturbations over the course of development (Waddington, 1942). Canalization in this context to the ability of an organism or population to produce the same phenotype regardless of variability of its environment (Waddington, 1959). Waddington used the metaphor of marbles rolling down ridges on a sloping landscape to illustrate canalization, with the marbles representing developmental pathways and the depth of the ridges representing the strength or weighting of canalized trajectories (Waddington, 1957). The deeper the ridges, the more resilient the development is to perturbations. Canalization acts to constrain phenotypic variation and stabilize development, leading to robustness and resilience (Flatt, 2005).

Carhart-Harris et al. (2022) repurpose the concept of canalization in application to ontogenetic psychological and neurobiological development, with special reference to mental health. Canalization, in referring to "phenotypic stabilization" (Waddington, 1959), is formally the opposite or inverse of "phenotypic plasticity" (Dewitt et al., 1998; Belsky and Pluess, 2013). Applied to cognition and mental health canalization refers to the strengthening and stabilization of cognitive or behavioral patterns over the course of development, such that they become more precise, dominant and entrenched over time (Debat & David, 2001). This progressive narrowing of potential pathways occurs through activity-dependent neuroplastic processes (e.g., associative or Hebbian learning) that selectively reinforce specific neural connections via selective synaptic potentiation or depression (Cicchetti & Tucker, 1994). Many disorders feature the pathological stabilization of particular cognitive-behavioral pathways, reflecting their shared reliance on (Hebbian) neuroplastic processes of canalization gone awry. Canalization, as the mechanism by which cognitive phenotypes are stabalized, is generally a normal part of adaptive development. It becomes psychopathological only in the case where it overshoots to stabalise phenotypes we identify as being pathological.

Many forms of psychopathology involve rigid, recurrent patterns of cognition, affect and behavior that persist in a canalized fashion, meaning they become stable and resilient, even in the face of conflicting external evidence or contingencies signaling the need for adaptation or change (Lewis & Todd, 2007). For example, the ruminative thought patterns in depression (Koster et al., 2011), ingrained substance cravings in addiction (Everitt & Robbins, 2005), obsessive intrusions in OCD (Pauls et al., 2014), body image distortions in eating disorders (Vocks et al., 2010), and delusional complexes in schizophrenia (Speechley et al., 2010) all demonstrate canalized cognitive-behavioral phenotypes that have become highly stable attractor states dominating the individual's mental and behavioral landscape.

These maladaptive patterns are theorized to develop in response to life adversity and associated distress or dysphoria; perhaps especially in early, i.e., the (human extended) critical periods for psychological and neurobiological development. The formation of canalized beliefs and behaviors is typically implicit. In line with the ‘free-energy principle’ (Friston, 2010), canalization can be understood as functioning to minimize uncertainty and associated negative emotional states. However, over the course of development, the implicitly chosen solution to reduce uncertainty (i.e., the defense mechanism) becomes automatic and habitualized. With excessive repetition and reinforcement over developmental time, patterns encoding the relevant thoughts and/or behaviors become reinforced to the detriment of healthy exploration and learning in the world; in the service of adaptability and behavioral and cognitive flexibility. Individuals become stuck in quasi-automatic loops that persist rigidly even when no longer functional. These loops are self-reinforcing and de-afferented from evidence from the external world (c.f., parasitic attractors).

Carhart-Harris et al. (2022) distinguish between two types of plasticity: an early form called "Temperature or Entropy Mediated Plasticity" (TEMP) that relates to increased variance, and canalization, where the latter is consistent with Hebbian plasticity and leads to a greater weighting of the encoded pattern or, in Bayesian terms, increased precision-weighting. They propose that "pathological" phenotypes develop via canalization mechanisms, as defensive responses to adversity. The canalized phenotypes and neural activity encoding them become entrenched and resistant to revision by external evidence informed updating. The depth of canalization determines the severity and treatment resistance of the pathology. The model states that canalization narrows the phenotypic state-space as ‘expertise in the pathology’ strengthens. TEMP, combined with psychological support and guidance, can counteract these processes by decreasing precision, akin to increasing temperature in the Ising model (Singleton et al, 2022), enabling a potentially therapeutic destabilization through which the dysfunctional stabilization could be weakened. Consistent with the phenomenon of simulated annealing (Kirkpatrick et al, 1983), TEMP functions to enable escape from canalized minima, and the exploration of alternative neural, cognitive and behavioral states. Like annealing in metallurgy and simulated annealing in computer science, the heating phase of TEMP followed by a cooling phase serves the recalibration or rebalancing of the attractor weights within the global state space. The canalization model proposes that this recalibration process causes a flatter or freer energy landscape in the service of improved mental health (Gómez-Emilsson, 2021; Hipolito et al, 2023). Specifically, the de-weighting of especially over-weighted (i.e., over-canalized) sub-states or attractors is key in this regard.

Canalization offers a testable mechanism through which the putative ‘p-factor’ of psychopathology comes about. However, as it is currently a theoretical sketch, it lacks specification as to how systems become pathologically canalized, and how psychotherapeutic mechanisms work and should be designed to counteract psychopathological canalization. Furthermore, the computational mechanisms and the relationship between canalization and psychopathology more generally remain unclear. In this paper we argue that adaptive narrative control theory can highlight the candidate computational mechanism underwriting psychopathological canalization. In so doing, we highlight how canalization as a putative common causal mechanism for psychopathology can be integrated into the field of computational psychiatry.

Section 2: Adaptive narrative control theory

The central idea of adaptive narrative control theory is that the contents of consciousness are shaped by sub-personal processes geared towards adaptive behavior rather than ‘veridical’ inference. Adaptive narrative control theory, stated simply, is the claim that pragmatic implications of attentional states are incorporated into (mental) action selection. In other words, the system controls the ‘narrative’ — the contents of conscious experience, or the multi-level web of predictions — not to veridically represent reality, but to drive adaptive behavior.

The capacity for adaptive narrative control is understood as a natural consequence of intelligent systems becoming more sophisticated: once a system can anticipate *how it will feel*, contingent on (mental) actions — and the kinds of reactive responses which will be potentiated by certain inputs and attentional states— it is able to select action policies to regulate affective states in the service of adaptive and functional behavior. For a person walking a tightrope, the strategy of ‘don’t look down’ allows them to control their affective responses and keep their composure to complete their goal. While strategies of this kind may enable short-term functionality, biased evidence selection can lead the model to fall out of step with the world as controlling affective states interferes with appropriate epistemic sampling.

We believe that narrative control pertains to beliefs subtending every spatiotemporal scale: agents like human beings are prone to directing their attention in service of adaptive behavior and affective and emotional regulation in order to achieve adaptive outcomes across all levels of belief formation. To give example on a short timescale, the boxer who (sub-personally) anticipates that eye contact with his opponent would lead to an affective response that would impair his performance may take ‘avoidant actions’ (e.g. avoiding eye contact with a fearsome opponent) in order to regulate his affective affective and energetic state be able to fight better. The same mechanism, we will argue, operates on longer timescales: for instnace, someone who has evidence suggestive of their partner’s infidelity— but also has a huge work deadline coming up— may attend away (both in their thoughts and actions) from the available evidence until they have more capacity to deal with it. In anticipating (sub-personally) that discovering infidelity would dysregulate their emotional state in a way that would disrupt their ability to complete their work, the person attends away from the evidence of infidelity. Both of these are cases of adaptive narrative control, and we contend that adaptive narrative control is happening all the time across multiple spatial scales: the system is *always* selecting the narrative that leads to (anticipated) adaptive outcomes — or the least expected free energy, or maximum reward, etc, depending on what kind of system it is. The brain is always taking into account potential consequences of holding certain personal level beliefs or perceptions. These examples involve suspending epistemic foraging through overt action, but we take systems with endogenous control of attention to be able to do this internally or ‘covertly’, through (for example) attending away from certain memories or beliefs.

While in the short-term biasing evidence selection to regulate affective states and achieve particular goals may be effective, it comes at a cost: the system loses out on valuable epistemic information, and develops a worse world model that it can use on longer timescales.From the brain’s perspective, then, it must strike a delicate balance between selecting evidence which leads to a counterfactually robust world-model over time, and selecting evidence which enables it to function effectively in the present and near future. We can anticipate how this evidence selection is context dependent: when a system is highly sensitive to perturbations in affective states, either dispositionally or due to acute stress, is likely to engage more in avoidant action to achieve narrative control and emotional stability.

We argue in this paper that psychopathology can be understood as a byproduct of the computational mechanisms of adaptive narrative control. Specifically, we propose (habitualised) avoidant mental action (or, equivalently, 'motivated inattention') as the key mechanism contributing to canalization in psychopathology. Avoidant mental action here is understood as biasing evidence selection for the purpose of regulating affective states. This is understood in many contexts to be adaptive — attenuation of certain inputs, memories, interoceptive or affective states is adaptive and functional in particular circumstances. While adaptive on short timescales, the perseverance avoidant mental action is understood to come with various costs: i) the biased selection of evidence leads to the agent's model of the world becoming more resistant to update via relevant new information; ii) attenuation of interoceptive inputs leads to poorer emotional granularity, which means the agent has impoverished models of the cause of stress and poorer means to resolve it. As such, we can understand the development of psychopathology to likely originate in adversity that leads to avoidant mental action as an affective regulation strategy, which is then habitualised, eventually resulting in the canalized phenotypes which characterize the psychopathology. Exactly how motivated inattention leads to psychopathology and belief rigidity depends on a number of factors, but different environments will motivate different patterns of inattention and, as a result, lead to distinct symptomatology and conditions.

Section 3: The active inference framework

The active inference framework is a specific articulation of Bayesian mechanics as applied to agentic systems (Ramstead et al, 2023). Bayesian mechanics refers to the equations of motion derived from the *free energy principle* (FEP) that describe how self-organizing dynamical systems evolve in time (Ramstead et al, 2023). The FEP is the mathematical statement that if something persists over time with a given structure, then it must entail or instantiate a statistical (generative) model of its environment. From these underpinnings in fundamental information theory, the active inference framework encompasses and extends upon previous inferential approaches to brain functionality, for instance building on Hermann von Helmholtz's theory that perception is a form of inference (Helmholtz, 1867). The active inference framework unifies perception, action, and cognition under a shared process of variational inference.

In the case of perception, it posits that the internal states of the brain or nervous system parameterise approximate Bayesian ‘beliefs’ or assumptions about states of the world and the body (Clark, 2013; Friston, 2010; Hohwy, 2013). Prior expectations about worldly regularities are combined with incoming sensory evidence to generate a posterior estimate — or Bayesian "best guess" of the hidden external causes of sensory input. At the sensory periphery, predictions track fast-moving and spatially local scales, while increasingly abstract representations are found at higher (i.e., deeper) levels of the cortical hierarchy (Kiebel et al., 2008). Perception, therefore, is understood as "perceptual inference," wherein predictions of the hidden causes or states in the world are iteratively updated to minimize incoming prediction errors (i.e., best explain sensory evidence). This notion of prediction arises from the FEP, notably that internal states can be shown to track external states for any persisting system (Da Costa et al, 2021). This relationship to prediction has led to an association of active inference with the wider umbrella concept of *predictive processing,* though it is worth noting that these are not formally identical (Kersten, 2023).

Active inference and the underlying free energy principle, while not theories of consciousness, have been identified as viable framework ‘for’ a theory of consciousness (Hohwy & Seth, 2020). A complete description of consciousness within the FEP literature is beyond the present scope, but an important central consensus is the notion that lived experience corresponds to the posterior inference. In the words of Hohwy (2013):

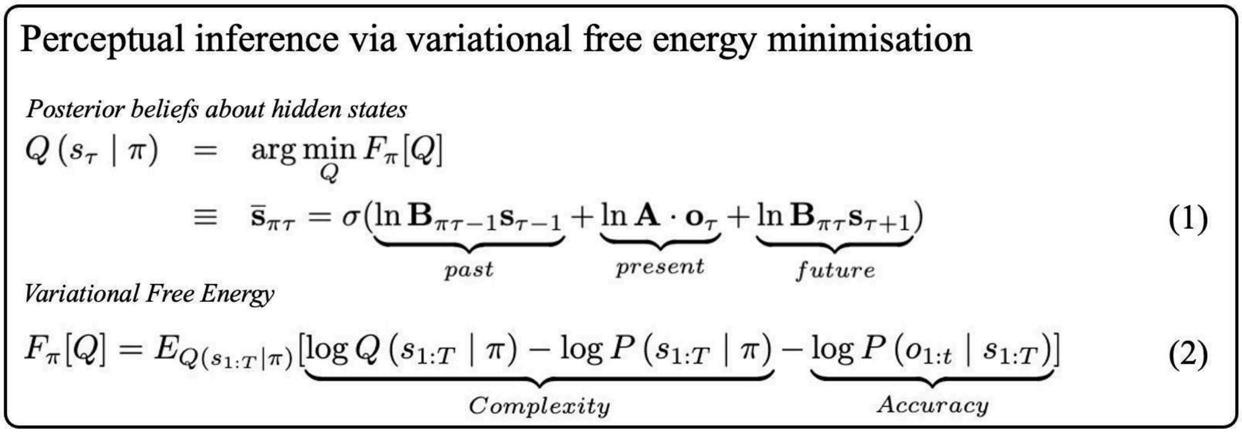
> Conscious perception is the upshot of unconscious perceptual inference. We are not consciously engaging in Bayesian updating of our priors in the light of new evidence, nor of the way sensory input is predicted and then attenuated. What is conscious is the result of the inference – the conclusion. That is, conscious perception is determined by the hypotheses about the world that best predicts input and thereby gets the highest posterior probability.

Active inference takes the Helmholtzian approach further by incorporating action selection into the inference process. This involves considering the system's expectations for maintaining homeostatic viability (Friston, 2010). Through this lens, active inference can be seen as a formalization of allostasis or predictive regulation of bodily states (Sterling, 2012). Essentially, the system opts for actions that it sub-personally "believes" will most likely lead to beneficial or ‘phenotype-congruent’ outcomes, i.e. those outcomes that promote survival and ongoing existence given the kind of creature that it is. Consequently, agents operating under active inference are frequently described as "self-evidencing", acting to increase the evidence for their own models (Hohwy, 2016), and the contents of phenomenology can be understood to reflect the allostatic imperatives of the agent (Deane, 2022; Ji et al, 2022; Nave et al, 2022).

Active inference extends the mechanics of perceptual inference to action by using a different approach to prediction error minimization: instead of solely updating the model to fit the world, active inference posits that the world (or sensory inputs assumed to be caused by states of the world) can be altered through action to conform to the model's expectations. Movement occurs through "systematic misrepresentation" (Wiese, 2017), where the prediction of the proprioceptive consequences of an action acts as a self-fulfilling prophecy (Adams et al., 2013; Brown et al., 2013). Accounts of consciousness within the active inference framework generally emphasize that it is self-modeling — through the modeling of the sensory consequences of action — that is central to understanding consciousness and subjectivity within the active inference framework (Deane, 2021; Friston, 2018).

The generative model is central to the process of inference underwriting both action and perception. A generative model delineates the probabilistic relationships (or Bayesian "beliefs") between observations and their underlying hidden states (the likelihood), and beliefs about how states evolve over time from some initial state (the prior). Prior beliefs pertain to the world before an observation and are updated to posterior beliefs, following an observation. This is known as Bayesian belief updating. The generative model is so-called because it generates the observable outcomes from unobservable states in the world (i.e., ‘hidden causes’). The active inference framework offers a conception of how agents select actions based on their expected sensory consequences. In this case, agents must reduce their uncertainty about states of the world, denoted by the variable *s*, based on their observations or "outcomes," denoted by the variable *o*. Generally, the agent selects actions—or sequences of actions called "policies," denoted by *π*—that elicit preferred outcomes and induce observations that resolve uncertainty about the hidden states (*s*). Posterior beliefs (about hidden states) should be parsimonious, minimizing the discrepancy between prior beliefs and observations while updating as little as possible (remaining close to the prior) (Mann et al, 2022; Whyte et al, 2022)

Inference at a given moment, i.e. perceptual inference, can be understood as the inversion of the generative model to infer the hidden causes of sensory signals. Direct computation of the posterior over hidden causes cannot be physically realized, so the organism is described as minimizing 'variational free energy.' This ensures that the Bayesian beliefs parameterised by internal states converge to posterior beliefs. Variational free energy can be expressed in terms of two penalties: the penalty for failing to explain the data (inaccuracy) and the penalty for overfitting (the complexity of the belief update). The variational free energy, as the sum of the inaccuracy and complexity, is minimized when accuracy and complexity are balanced against each other. The 'best guess' strikes this balance to reach a posterior belief about hidden causes.



Perceptual inference via variational free energy minimisation

Posterior beliefs about hidden states

$$\begin{aligned} Q\left(s_\tau \mid \pi\right) &= \arg\min_Q F_\pi[Q] \\ &\equiv \bar{\mathbf{s}}_{\pi\tau} = \sigma\big(\underbrace{\ln \mathbf{B}_{\pi\tau-1}\mathbf{s}_{\tau-1}}_{past} + \underbrace{\ln \mathbf{A}\cdot\mathbf{o}_\tau}_{present} + \underbrace{\ln \mathbf{B}_{\pi\tau}\mathbf{s}_{\tau+1}}_{future}\big) \end{aligned} \qquad (1)$$

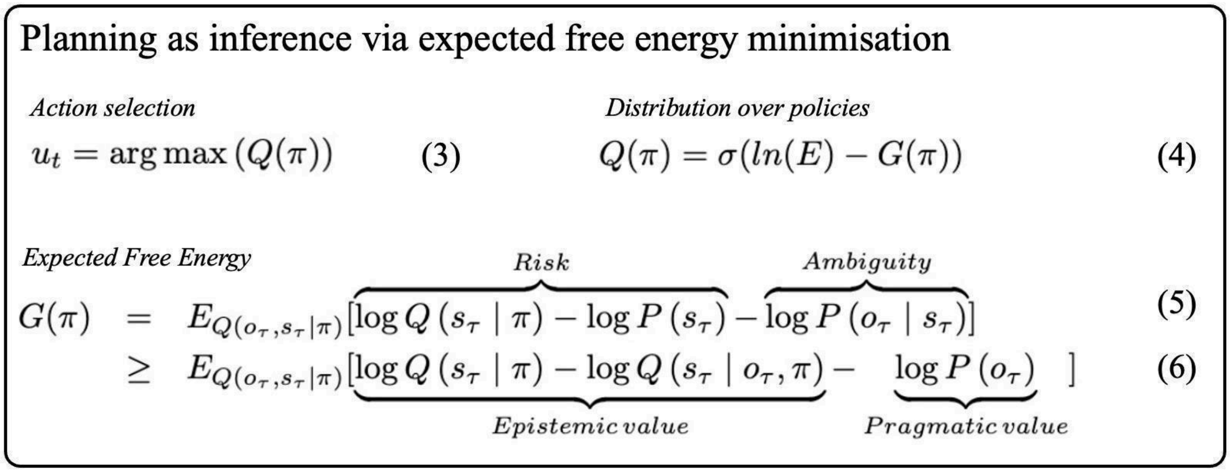
Variational Free Energy

$$F_\pi[Q] = E_{Q(s_{1:T}\mid\pi)}[\underbrace{\log Q\left(s_{1:T}\mid\pi\right) - \log P\left(s_{1:T}\mid\pi\right)}_{Complexity} - \underbrace{\log P\left(o_{1:t}\mid s_{1:T}\right)}_{Accuracy}] \qquad (2)$$

Figure 1: Equations of perceptual inference derived from the FEP. These equations illustrate the process of perception as optimizing a posterior distribution over hidden states s_τ at time τ given the policy π, denoted $Q(s_\tau \mid \pi)$ (or $s_{\pi\tau}$ when expressed in terms of its sufficient statistics as in (1)). This optimisation is usually cast as a gradient descent on the variational free energy (VFE), under the assumption that the generative model takes the form of a discrete, partially observable Markov decision process (POMDP). It states that evidence based on transition beliefs, B, and the likelihood of sensory data, A, are combined using the softmax function σ to determine the posterior. The VFE, equation (2), is a bound on self-information or the (negative log) evidence or marginal likelihood of any given observation. This means that minimizing VEB is equivalent to maximizing the evidence for a generative model of sensory observations. Because self information scores the surprisal (a.k.a., surprise) of the observation under the generative model, it can also be read as a prediction error. This means that maximizing marginal likelihood or model evidence is the same as minimizing surprise or prediction error. The variational free energy can be decomposed in a

few equivalent ways, here expressed in terms of the complexity and accuracy terms, demonstrating that VFE minimizing beliefs are those with an optimal balance of low complexity and high accuracy.

While variational free energy pertains to inferring hidden states in the present, *expected free energy* concerns the free energy the agent expects under a particular action policy. In perceptual inference, the system infers the state of the world to be the state with the least variational free energy. In active inference, actions are selected on the basis of expectations about transitions between states contingent on action. The system samples from a distribution over potential actions, where actions with the least expected free energy have the highest probability of selection. Conceptually, minimizing expected free energy can be understood as selecting actions that maximize reward and information gain, since the expected free energy is a measure that integrates both the epistemic value (or information gain) and the pragmatic value (or goal-directed utility) associated with potential actions: see equation (6) (Friston et al., 2015). In other words, expected free energy combines the organism's desire to minimize uncertainty about the world (i.e., reducing prediction errors) with its drive to achieve goals or fulfill prior preferences (Friston et al., 2016). By minimizing expected free energy, organisms can balance the trade-off between exploration (i.e., information seeking) and exploitation (i.e., goal seeking) in an uncertain environment. Organisms can therefore optimize their actions to both gain valuable information about the world and fulfill their preferences or goals. Active inference nuances the notion of goals (and rewards) in an important way: goals are simply outcomes that are characteristic of the agent in question, as scored by the prior probability the agent will encounter certain outcomes. This means actions are selected under prior beliefs that can be read as prior preferences for outcomes that characterize the agent.



Planning as inference via expected free energy minimisation

Action selection

$$u_t = \arg\max\left(Q(\pi)\right) \qquad (3)$$

Distribution over policies

$$Q(\pi) = \sigma(ln(E) - G(\pi)) \qquad (4)$$

Expected Free Energy

$$\begin{aligned} G(\pi) &= E_{Q(o_\tau, s_\tau \mid \pi)}[\overbrace{\log Q\left(s_\tau \mid \pi\right) - \log P\left(s_\tau\right)}^{Risk} - \overbrace{\log P\left(o_\tau \mid s_\tau\right)}^{Ambiguity}] && (5) \\ &\geq E_{Q(o_\tau, s_\tau \mid \pi)}[\underbrace{\log Q\left(s_\tau \mid \pi\right) - \log Q\left(s_\tau \mid o_\tau, \pi\right)}_{Epistemic\,value} - \underbrace{\log P\left(o_\tau\right)}_{Pragmatic\,value}] && (6) \end{aligned}$$

Figure 2: Equations of active planning as inference derived from the FEP. These equations illustrate the process of selecting an action u_τ at time τ given the probability distribution over policies $Q(\pi)$. Equation (4) shows that high probability policies are those with high prior probability E and low expected free energy G(π). The expected free energy scores the negative log likelihood of each policy π. It can be expressed in terms of the expected risk and ambiguity (5), or in terms of the expected epistemic value and pragmatic value (6), demonstrating that — by selecting actions that minimize G(π) — the agent strikes a balance between explorative and exploitative behavior.

Active inference has been characterized as a formalization of the concept of allostasis—'stability through change'—which is the predictive regulation of bodily states through action—such as seeking shade on a hot day. Systems with expectations of how states will change over time, contingent on actions, can select actions and policies that avoid dyshomeostatic outcomes before they arise, and tolerate perturbations to homeostatic setpoints when better outcomes are expected over longer timescales, including tolerating worse outcomes in the short term in favor of better outcomes on longer timescales (c.f., ‘deferred gratification’). Active inference formalizes allostasis as a problem of inference ('planning as inference') (Kaplan & Friston, 2018), where allostasis includes effectively budgeting resources for action. When allostatic mechanisms are overused or maintained over extended periods — when stress isn’t effectively moderated through action — this can lead to wear and tear known as allostatic load (Arnaldo et al., 2022). Markers of allostatic load include imbalances in primary mediators like hormones as well as secondary outcomes like inflammatory biomarkers. Ultimately, allostatic load (i.e., ‘stress’) can increase vulnerability to pathologies like depression as the body's systems fail to effectively regulate in response to stressors (Arnaldo et al., 2022). Allostatic ‘overload’ (Guidi et al, 2020) can be likened to the construct of ‘decompensation’ - i.e., a collapse or breakdown of compensatory mechanisms. This construct is useful as, like canalization, it traverses both the physiological and psychological domains. The constructs of the ‘pivotal mental states’ (Brouwer et al, 2021) and ‘quantum change’ (Miller et al, 2001) are also relevant here.

In summary, active inference casts action as a process of inference where action selection can be understood as the selection of transitions between hidden states expected to have the least expected free energy, which are therefore most likely to lead to self-evidencing (phenotype-congruent, survival-promoting) outcomes. By integrating perception, action, and cognition under a common process of inference, active inference allows agents to navigate complex environments and adapt their behavior to maximize their chances of survival and overall well-being.[1]

Attention, mental action, and affective inference

In active inference (and predictive processing more generally) attention is conceptualized in terms of the precision, or inverse variance, of the likelihood mapping (Feldman & Friston, 2010). Precision in active inference represents a second-order confidence in, or weighting of, some belief captured by a parameter of the generative model — in other words; a belief about a belief. The likelihood precision is the confidence in the probabilistic relationship between hidden states and observations (i.e. the likelihood mapping), and can be understood as the degree to which the system believes its observations are or will be generated precisely by hidden states. This form of precision underlies attentional modulation in perception; for example, focusing on a specific instrument in a musical piece increases the agent's confidence in determining the cause, such as identifying which instrument is being played. In other words, attending to the instrument mechanistically entails increasing the gain in one part of feature space. In other words, attending to a particular instrument in this case mechanistically entails increasing the gain on part of the feature space. Likelihood precision — often denoted by ζ — acts as a form of gain control on the sensory signal i.e., in Bayesian terms, it is the precision of the prediction errors that is positively modulated via attention.

[1] Strictly speaking, the free energy principle turns this on its head and says that anything that survives — in some characteristic or preferred states — can be described as minimizing variational and expected free energy; and can therefore be described teleologically in terms of active inference (Constant, A, 2017)

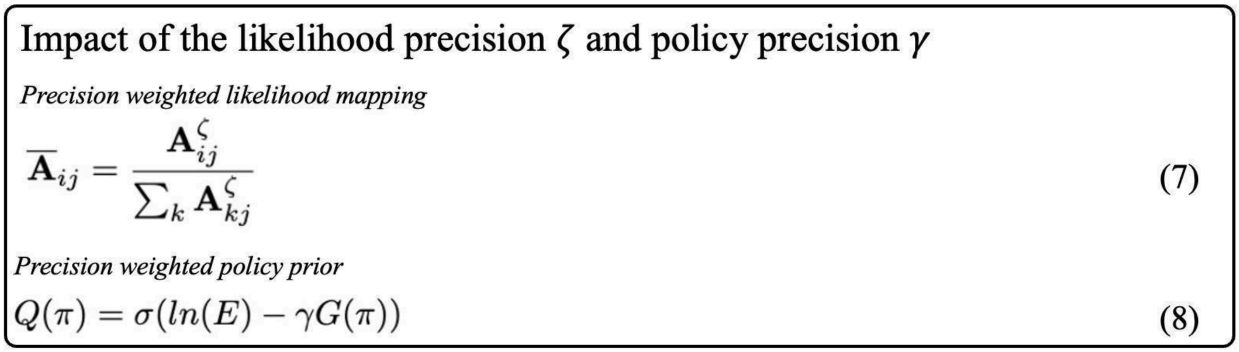
The counterpart of attention is sensory attenuation, which refers to the top-down attenuation of incoming information, both from the body (interoception) and the external environment (exteroception). Attenuation, therefore, is a reduction of the likelihood precision — e.g., downweighting of prediction error signals —which plays a crucial explanatory role in active inference in relation to the mechanics of action. Attenuation of proprioceptive inputs, or 'physiological sensory attenuation' (Palmer et al., 2016), is essential for action, as the system ignores evidence for the current state (e.g., my arm is not moving) and allows prior beliefs (e.g., I am lifting my arm) to be realized and thereby initiate movement. Selectively attenuating (lowering precision) on either interoceptive or exteroceptive inputs enables the system to filter out irrelevant inputs, such as reafferent inputs resulting from self-generated actions. Sensory attenuation of the anticipated consequences of action is thought to account for the inability to tickle oneself (Blakemore et al., 2000) but this is just one of a multitude of possible examples, as sensory attenuation is hypothesized to be a major, permissive and dominant feature of normal waking consciousness (Palmer et al., 2016); ranging from saccadic suppression during eye movements to gymnastics. Attenuation is the flipside of attention, and is the natural byproduct of attending to a particular stimulus: attending to one thing (increasing the gain on one stream of information) necessitates attenuating inputs from another thing (decreasing the gain on a stream of information).

The process of likelihood precision modulation — in the context of sensory attenuation for motor action — is generally considered to be an automatic process. However, the mechanics of precision estimation, furnished by the active inference framework, enable the formalization of precision modulation as a form of mental action (Limanowski & Friston, 2018, Sandved-Smith et al., 2021). This means that endogenous and exogenous shifts in attention can be modeled as a process of expected free energy minimisation over states of attention which drive the deployment of precision. In active inference, the system can select policies to transition between hidden states to produce certain (expected) outcomes. Mental action is understood to work in much the same way, however in this case the transitions between states are not between the (inferred) states of the world, rather state transitions track transitions between attentional states or predicted precision (Corbetta & Shulman, 2002). The system does not require direct access to precisions to achieve this, only expectations for outcomes contingent on actions — for example, a mental action of shifting attention may come with the sensory expectation that one stimulus becomes more highly salient and precise than another.

The active inference framework is therefore equipped to incorporate concepts from Thomas Metzinger's self-modeling theory of subjectivity. The principle of transparency, as introduced by Metzinger, refers to the idea that we usually do not perceive our own cognitive processes, but rather the content they generate. For example, when we see a tree, the experience of attending to the tree is usually 'transparent', that is, we don't perceive the dynamics of belief updating that underpin this visual experience, but simply experience the tree (Metzinger, 2003). This is synonymous with saying that the inferential processes are implicit or unconscious. The dynamics of likelihood precision (which capture the attentional dynamics) is an example of a parameter of the generative model that factors into the perceptual process (i.e. the posterior estimation depends on it), but that does not feature as content in perceptual inference. In contrast, the concept of opacity represents a situation when we perceive the cognitive processes themselves, rather than just the content they make available, i.e., they are explicit. When cognitive processes 'become opaque', we can perceive, for instance, the shifts in our attentional states *per se*. In this situation, the likelihood precision is the evidence for a second-order posterior inference, which gives rise to an experience about the processes of my perception. In general, it is this notion of opacity that allows us to, under certain conditions, metacognitively register our thoughts, perceptions, or feelings as constructs of our mind rather than direct reflections of reality (Metzinger, 2003). Under active inference, only those states that can be inferred are amenable to the action selection process, therefore, it is this notion of opacity that makes mental contents amenable to mental action and modulation (Metzinger, 2017: Limanowski & Friston, 2018, Sandved-Smith et al., 2021).

To model agents such as ourselves, who are capable of being aware of their cognitive processes and modulating them to some degree, requires a form of hierarchical inference that has been referred to as *parametric depth* (Sandved-Smith et al., 2021). This architecture enables the agent to form second order inferences about the parameters of the generative model, which would otherwise remain transparent. This extends, in principle, to a number of parameters that the agent can form inferences about, namely the various precision estimates of the generative model.

Another notion of precision concerns the precision over the action model, defined in terms of a precision (i.e., inverse temperature) parameter over the expected free energy (Hesp et al, 2021). This estimate scores how confident the system is in its prior beliefs about action. Action selection in active inference comprises three essential components: 1) the prior over policies, equated with 'habit'; 2) the expected free energy estimates associated with candidate action policies; and 3) the precision over the action model, generally specified by a gamma parameter and understood as the system's confidence in its own action selection (see equation (8)). The gamma parameter scales the expected free energy estimates, such that high confidence results in the system selecting policies based on its estimates of which policies have the least expected free energy, while low confidence (lower values of gamma) means the system has lower confidence in its expected free energy estimates, and policy selection instead is dominated by the prior over policies.



Impact of the likelihood precision ζ and policy precision γ

Precision weighted likelihood mapping

$$\overline{\mathbf{A}}_{ij} = \frac{\mathbf{A}_{ij}^{\zeta}}{\sum_k \mathbf{A}_{kj}^{\zeta}} \tag{7}$$

Precision weighted policy prior

$$Q(\pi) = \sigma(ln(E) - \gamma G(\pi)) \tag{8}$$

Figure 3: Precision weighted perceptual and action models. These equations enrich the expressions of the likelihood mapping A and policy distribution $Q(\pi)$ with second order beliefs about the precision of the associated parameters. In (7) the precision weighted likelihood mapping is obtained by exponentiating each element in the *i*th row and *j*th column of A by ζ and normalizing (see Parr et al. 2018). Notice that as the precision tends to zero, the sensory evidence term in (1) tends to zero. In (8) the distribution over policies is modulated by the model precision γ, the basis of affective state inference. Notice that as γ tends to zero (low affect), the policy selection biases towards habitual priors E. Note that the first equality is one of several ways of parameterising precision. In this instance, it appeals to a Gibbs distribution over the likelihood parameters.

In affective inference, the system not only possesses an estimate of action model confidence, described as 'implicitly metacognitive,' but another layer of the generative model can be added so that the confidence estimate itself becomes a hidden state that the system infers using the same mechanics of perceptual inference. In this case, the system represents itself in a specific state, for example, confident, or not, in its action model, which Hesp et al (2021) argue corresponds to positive and negative valence. That affective states are opaque; that is, the system has expectations not only for states of the world contingent on actions, but also expectations for its own action model confidence, or ‘allostatic control’, contingent on actions and mental actions. This kind of parametrically deep generative model is key to the account presented in the next section.

This section has considered the mechanisms of attention, mental action, and affective inference under the active inference framework. It is worth summarizing briefly the distinct notions of precision introduced, to clarify how different precision terms relate to different parts of the generative model (Hesp et al, 2021; Sandved-Smith et al, 2021; Smith et al, 2022). 'Sensory precision' refers to the confidence assigned to sensory observations, serving as a measure of the reliability or trustworthiness of these inputs in forming accurate perceptions of the environment. Typically in the active inference literature this is encoded as a weighting, ζ, of the likelihood or ‘A matrices’. Finally, policy (or model) precision, γ, refers to the system's confidence in actions or, in other words, confidence in the action model, and is understood as the basis of affective inference. Importantly, a system able to model policy precision explicitly, can infer "what would my policy precision be contingent on different (mental) action policies," which corresponds to "how would I feel if I were to take this (mental) action."

Section 4: The mechanism of motivated inattention

The overview of active inference lays the foundation for understanding how canalization of psychopathological traits occurs as a natural consequence of the computational mechanisms underwriting action and perception in sufficiently deep systems. Specifically, we propose ‘avoidant mental action’ (or ‘motivated inattention’) as the mechanism leading to psychopathology and canalization. The central claim here is that systems with sufficient parametric depth, such that their affective and attentional states are opaque, will be driven to select certain mental actions in order to regulate affective states. While regulating affective states in this way can be part of adaptive action selection in the short-term, it involves biasing evidence sampling in a way which can bring the generative model out of alignment with environmental contingencies, which can be maladaptive and lead to psychopathology and decreased subjective well-being on longer timescales.

A system that models which affective state it will find itself in, contingent on mental actions, can attend to stimuli on this basis to bring about particular affective states that are expected to provide functionality in a given context. This includes the selection of mental actions which attenuate inputs that are expected to destabilize or bring about undesirable affective states. This mechanism, which can be understood as a kind of 'motivated inattention' or 'avoidant mental action', bears striking resemblance to the Freudian concept of repression — it is a psychological defense mechanism where unpleasant thoughts, feelings or memories are systematically attenuated and as such do not register in consciousness.

The explicit modeling of (hidden) affective states is achieved by performing second-order inference on the "model precision", that is, the system's confidence in its action model (encoded in the gamma parameter). As a result, the system will have a predictive model of its model precision, and expectations for transitions between model precision estimates contingent on actions—intuitively, "how will I 'feel' if I do that?". In action planning, this predictive model is compared with the agent's preference model (the states the agent expects to occupy). Actions can then be selected in order to minimize the divergence between the prediction and preference models for affective states.

With this architecture in place, an agent can take actions to realize certain affective states: a specific action may have the expected outcome of increasing confidence in the action model, compared with another action, which may have the expected outcome of decreasing confidence in the action model (manifesting phenomenologically as negative affective valence). Just as actions are selected on the basis of expectations about transitions among hidden states of the world, mental actions are understood to be selected based on expectations of how attentional states (precision estimations) will transition contingent on the mental action. Consequently, evidence that is likely to reduce confidence in the action model can be attenuated through mental action selection, a form of 'motivated inattention' or 'avoidant mental action'. The upshot is that, in biasing evidence selection through selective attenuation in this way, the system can maintain high confidence in the action model — at least in the short-term. For example, we might expect that attenuation of evidence or memories of failures at flying a plane might be acutely beneficial for a nervous pilot during take-off, in order to avoid destabilizing their emotional state and impairing their flying ability. This maintenance of a particular affective state, while enabling short-term functionality, comes at the cost of the system not incorporating important information into its generative model.[2]

A natural objection to this account is that it seems essential for an agent to *not* ignore crucial information that might lead to decreased confidence in their action model, as adjusting this confidence is vital for selecting context-appropriate action. For instance, if a specific observation would encourage the agent to be more cautious in their behavior, or prompt them to seek epistemic actions before satisfying prior preferences, it might be detrimental for the agent to not update their confidence based on these observations to enable adaptive action selection.

We propose that this mechanism is general, and frequently adaptive. Attenuating certain inputs (indeed the majority of inputs) is essential for functionality. Furthermore, there are many contexts where affective regulation—through motivated and selective inattention—is essential for the agent to be able to function effectively. Attenuating certain memories, interoceptive inputs, or sensory evidence may be beneficial in a wide variety of contexts. This proposal is motivated by the function of valence as articulated by the gamma parameter in equation (8), which modulates the relative impact of habits (E) in action selection. A sophisticated agent capable of modeling this consequence of their future affective states will be able to, for instance, attenuate sensory evidence that might decrease valence and result in suboptimal habitual action. For example, in order to avoid anxiety induced disruptions in motor control (Harris et al., 2023), a public speaker might ignore a negative text message before getting on stage in order to retain a fluid delivery instead of reverting to their nervous habit of disfluencies (“ums” and “uhs”).

[2] We might expect that in most instances, biased evidence selection would be associated with avoiding low valence affective states, and promoting high precision on the action model, in accordance with the view that an ‘optimism bias’ may be adaptive (Sharot, 2011). But this is not always the case: there are various instances where the selected policy might involve concealing (through suppression) a high valence affective state to someone else (such as a concealing elation at a hand in a game of poker), or staying on task requires the system attenuating inputs leading to high valence.

However, this mechanism (of avoidant mental action) can become pathological if it becomes habitualised or ‘canalized’. Experience-dependent (i.e., Hebbian) learning, the mechanism of neuroplasticity proposed to underlie canalization (Carhart-Harris et al, 2022), serves to allow repeated actions to become habitual and automatic. In other words, it supports rapid and efficient execution of the habitualised action (Hebb, 1949; Dayan & Berridge, 2014). In repeating certain forms of avoidant mental action, certain patterns of avoidance and attenuation become habitualised, meaning that the relaxation of these habitualised mental action routines becomes difficult or psychologically impossible. Computationally this would correspond to a strong prior belief over mental policies encoded in the associated E matrix. The model is then biased towards particular observations and against others, and in being less likely to incorporate the available evidence, becoming canalized and rigid.

This account furnishes a neurocomputational articulation of the Freudian notion of the ‘dynamic unconscious’ i.e., that which is withheld from awareness (i.e., ‘repressed’) through (unconscious) evasive mental action that attenuates evidence expected to decrease confidence in the action model. Unconscious contents, by being avoided, do ultimately determine the contents of conscious experience. They do so by influencing which assumptions and observations are weighted and thereby affect the resultant posterior, as a kind of “self-fulfilling sampling”. We consider this in more detail in section 6. In the next section, we consider how this mechanism leads to low subjective well-being and psychopathology.

Section 5: How motivated inattention leads to low subjective well-being

In the previous section, we provided an overview of how avoidant mental actions can become habitualized, leading to the "repression" of unpleasant emotions, memories or thoughts. In this section, we provide an overview of how the computational mechanisms of motivated and avoidant mental action can lead to decreased subjective well-being and psychopathology through: 1) preventing the appropriate update of the generative model (and prior preferences in particular) in light of available evidence, and 2) reducing emotional granularity and increasing allostatic load as a result.

We have argued that in order to promote or maintain certain affective states, the system engages in biased evidence selection through the attenuation of specific sensory evidence. One result of biasing the selection of evidence, in this way, is that the generative model is not accurately updated to reflect the dynamics of hidden causes in the world. For example, take a case where the system may select avoidant mental actions in order to regulate affective states and maintain functionality in a given context — such as attenuating evidence of failure to reach a certain goal state, or a worse than expected rate of goal realization. Attenuation of evidence of an unwanted or undesirable outcome, in order to regulate an affective state, leads the system to continue to pursue action policies that should be adjusted or recalibrated in terms of what is actually realizable — for example, ignoring evidence that a particular career path is unlikely to work out could lead to persistent frustration as the goal is not updated. The consequence of pursuing policies that are unlikely to be realizable is the accumulation of prediction error on longer timescales, leading to decreased confidence in the action model (and associated decrements in subjective well-being). Furthermore, the biased sampling of evidence, driven by affective state regulation, will lead to beliefs about states of the world being updated in a way which ignores potentially relevant information about the environment and environmental volatility. For example, someone in denial about a recent financial loss may continue to spend money, and so is out of step with the current environmental contingencies. There are many examples in psychiatry that one can call upon to illustrate the basic mechanism on offer. The crosscutting theme is that the capacity to attenuate or ignore evidence — requisite for action — can lead to actions that fail to sample evidence that need to be incorporated to update beliefs to reflect the world and provide realistic expectations of environmental contingencies and prior preferences. This leads to a canalisation of prior beliefs about overt and covert action that can become suboptimal in a changing world.

The second prediction is that motivated inattention leads to reduced emotional granularity, and increased allostatic load as a consequence. Within active inference, emotions are often understood in terms of interoceptive inference, incorporating multi-modal evidence to infer what bodily states ‘mean’ for action. Lisa Feldman-Barrett has expanded this approach into the ‘theory of constructed emotion’ (Barrett, 2017). According to this theory, emotions result from the integration of prior expectations with interoceptive sensory signals to generate the best guess of the hidden causes underlying these signals. For instance, a rapid heart rate and increased skin conductance can signify different things depending on the context, perhaps anxiety or sexual arousal, and so broader contextual cues are required to infer the cause of the physiological response, and to 'construct' the appropriate emotion. Physiological reactions may prepare the body for action within a given context, and emotions are thus "constructed" or "inferred" through the recruitment of predictive models to explain afferent interoceptive signals (Siegel et al., 2018; Wager et al., 2015).

Granularity plays a crucial role in emotional experience, relating to emotional complexity, diversity (Kang & Shaver, 2004; Quoidbach et al., 2014), and awareness (Lane & Smith, 2021; Smith & Lane, 2015; Smith, Killgore, & Lane, 2018). Alexithymia, for example, is associated with lower emotional granularity (Bagby et al., 1994; Lane et al., 2021; Maroti et al., 2018; Trevisan et al., 2019). Individuals with poor emotional granularity tend to use coarse-grained categorizations of their emotional states, such as 'good' or 'bad', tantamount to the 'black and white thinking' that is particularly characteristic of emotional unstable personality disorder (Kernberg, 1967). In contrast, individuals with highly granular generative models of their own affective states can identify specific emotion categories; e.g., joyful, angry (Smith, Lane, et al., 2019; Smith, Parr, et al., 2019). Low granularity constrains the information available to guide adaptive action.

Smith et al. (2022) unpack the novel computational perspective on emotional granularity provided by the active inference framework. Recall, within active inference, the brain maintains an internal model of the environment that is used to minimize prediction errors and guide adaptive behavior. Granularity can be represented as the size and specificity of the state space containing hypotheses or emotion concepts within this generative model (Smith et al., 2022). Individuals with high granularity have access to a large set of fine-grained emotion concepts for interpreting experiences (e.g., many distinct categories of negative affect like anger, sadness, guilt; Smith & Lane, 2015). This corresponds to a model with greater specificity in the state space. In contrast, those with low granularity utilize broad undifferentiated categories (e.g., simply “positive” or “negative” emotions), equivalent to a model with few possible states (Smith et al., 2022).

Smith et al. propose that higher granularity provides a more nuanced predictive model of the environment, allowing more precise inferences about the causes underlying sensory input. Access to multiple specific emotion concepts affords detailed information to guide perception, learning, decision-making and, crucially, mental action (Smith & Lane, 2016; Smith et al., 2022). For instance, the ability to differentiate “anger” from “disgust” supports selecting situation-appropriate actions. In contrast, reliance on coarse emotion categories limits differentiating information for effective regulation and choice (Smith et al., 2022). Indeed, higher emotional granularity has been associated with greater emotional complexity, emotional awareness, and adaptive social functioning (Kashdan et al., 2015; Quoidbach et al., 2014; Smith & Lane, 2016). Another relevant construct here is ‘tolerance of uncertainty’ (Patel & Hancock, 2023).

Granularity of affective responses is critical to avoid overgeneralization. Agents better able to identify their affective responses can generate more precise actions in response to various stimuli, avoiding overgeneralized low allostatic control and self-efficacy, which may lead to depression and low mood (Stephan et al, 2016). The same point applies in the other direction: agents better able to identify their preferences can find other things that share desirable (‘ego-syntonic’) properties.

How, then, does motivated inattention lead to decreased emotional granularity? Motivated inattention as described in this paper leads to the selective attenuation of inputs the system predicts will dysregulate affective states in a way which will disrupt or impair functional behavior in a given context. Given the centrality of interoceptive inputs in emotion formation, attenuation of interoceptive inputs is particularly likely. When habitualised, the agent never receives the fine-grained interoceptive inputs to form, or maintain, the fine-grained emotion concepts (and associated granular action policies) to resolve the source of the stress. Without the need to account for the nuanced signals, the agent’s generative model is likely to reduce its complexity: e.g., by eliminating redundant aspects of the generative model via bayesian model reduction (Friston et al., 2017). Furthermore, biasing evidence selection through attenuation of relevant interoceptive inputs can plausibly lead to mischaracterisation of emotions, such as misattribution of the stress to the hypotheses available to the agent under biased evidence selection. This could lead to either emotions too general to be rectified via actions (e.g. experiencing as frustration what would more properly categorized as anger), or actions that are unlikely to remedy the cause of the stress.

Failure to translate stress responses into adaptive and allostatic action, due to their causes being specified at a level of granularity that is too coarse to prescribe specific remedial action, leads to a build up in the ‘allostatic load’. Stress responses cannot be converted into stress alleviating action. As such, persistent stress leads the system to infer itself as in a position of low 'allostatic self-efficacy' — whereby it infers low confidence in its own ability to effectively budget energetic resources for the purpose of regulating bodily states. Stephan et al (2016) argue that this leads to depression and chronic fatigue, as the system learns to have low confidence in its own domain-general allostatic control (Ramstead et al 2020; Stephan et al 2016).

This section has outlined how avoidant mental action, akin to repression, can lead to low allostatic control. In summary, attenuation of interoceptive inputs, sensory evidence and memories through mental action leads the system to have poor emotional granularity. Poor emotional granularity results in increased allostatic load, as the system is not able to find the appropriate actions to ameliorate the stress response. Chronic increases in allostatic load lead the system to infer itself as in a state of low allostatic control and low self-efficacy, leading to lowered subjective well-being and depression. Here, we focus on the mechanisms of depression, but the specific manifestations of psychopathology caused by motivated inattention will vary according to the patterns of attention required to function in particular environmental conditions, and how these contribute to narrativization of experience.

Section 6: The computational unconscious

What is the fate of the attenuated and unattended contents in this picture? In being selectively suspended from awareness due to a subpersonal mechanism, the current account can be understood as a computational analogue to Freud’s conception of the dynamic unconscious.

The unconscious mind, a central concept in Freud's (2012) theory, refers to a primitive mental system, sometimes called the ‘system unconscious’ in Freud’s earliest work and later called the ‘it’ or ‘id’ in the standard English translation. Freud distinguished between the conscious mind, which encompasses thoughts and feelings we are aware of in the present moment; the preconscious, containing thoughts and feelings that are not immediately accessible but can be easily retrieved; and the unconscious, comprising thoughts, feelings, and memories and other mental content that are not readily accessible either because they were never conscious or because the content is kept hidden e.g., due to its potentially distressing and conflictual nature — what Freud called its ‘ego-dystonic’ nature — meaning ‘offensive to one’s self schema’ (Freud, 2014). While the 'descriptive unconscious' does not differentiate between preconscious and unconscious (Freud, 1912; 1922; Laplanche & Pontalis, 2018), the term 'dynamic unconscious' specifically refers to mental contents that are defensively removed from consciousness through the mechanism of repression.

The dynamic unconscious includes anything excluded from awareness via a defense mechanism such as repression. Freud thought that we must assume the existence of the unconscious via its effects, as we lack direct access to it. Examples of unconscious contents might include unwanted desires, traumatic memories, or threats to self-esteem. Freud described the dynamic unconscious as a repository of unresolved conflicts, repressed desires, and traumatic memories. Repression — ‘motivated ignorance or forgetting’— is thought to be the fundamental concept in Freudian psychoanalysis; it refers to the psychological mechanism through which distressing thoughts, feelings, and memories are unconsciously pushed out of conscious awareness to protect the individual from experiencing anxiety or emotional pain (Freud, 1915). According to Freud, repression serves as a defense mechanism that helps individuals maintain a sense of psychological equilibrium and avoid confronting uncomfortable or threatening aspects of their inner lives. Repressed mental contents, although hidden from conscious awareness, continue to influence the individual's behavior, emotions, and thought processes, often manifesting as maladaptive behavioral or mental patterns, or even the symptoms of mental illness (Freud, 1900). Freudian psychoanalytic therapy aims to uncover and resolve repressed conflicts and memories, thereby allowing the individual to achieve greater self-understanding and psychological well-being (Freud, 1912). In his view, while repression of certain content is normal in healthy human functioning, the mechanism is also considered dysfunctional and causative of mental health symptoms to the degree that the unconscious mental content seeks expression in somatic, cognitive or behavioral symptoms that cannot be worked through consciously as the subject is not aware of their unconscious source.

One distinct advantage of our proposed account is the potential to integrate the unconscious into computational psychiatry — the emerging interdisciplinary field that develops computational models of the underlying mechanisms of mental disorders. The concept of the dynamic unconscious remains a foundational concept in psychotherapy and psychoanalysis to this day. However —with exceptions (Carhart-Harris & Friston, 2010; Fotopoulou, Pfaff & Conway, 2012; Solms, 2017; Smith & Lane, 2016) — its influence remains limited in philosophy, neuroscience, and computational psychiatry. The exclusion of unconscious processes from these models could potentially limit the scope of investigation and overlook important aspects of mental functioning that traditional psychoanalytic theories have sought to address (Leichsenring & Rabung, 2011). Computational specification of avoidant mental action opens the door to thinking about how defense mechanisms such as repression can lead to decreased subjective well-being and psychopathology. A valuable follow on to this paper would be the implementation of the computational model presented here in order to test the hypotheses by simulating the attentional dynamics and the ensuing changes to the agent's behavior and subjective well-being.

Section 7: From sophistication to psychopathology and back again: mechanisms of psychotherapy and improved subjective well-being

This paper has presented a view of psychopathology as largely underwritten by a common mechanism which engenders rigid belief structures associated with canalization. This mechanism is thought to be an adaptive response in systems which become sophisticated enough to be able to anticipate (and therefore control) their own mental states and behavioral responses under particular conditions. Simpler systems lacking this capacity would, therefore, lack the associated psychopathologies. With this view in mind, the next section considers the route back: what are the factors that underlie the journey into wellness and improved subjective well-being?

We have postulated avoidant (mental) action as a core mechanism of psychopathology, and we now propose that the inverse of this account may provide a common computational mechanism across psychotherapeutic interventions and improvements to subjective well-being.

Earlier, we argued that psychopathology and low subjective well-being is likely downstream of avoidant mental actions that become habitualised and canalized into recalcitrant and rigid prior beliefs that bias inference, lock the agent into inflexible modes of action and perception, in a way that (1) precludes alignment between the generative model and the changing world and (2) blunts emotional granularity, leading to increased allostatic load and low (inferred) allostatic control. Under this formulation , we propose candidate mechanisms of psychotherapy. We take these to be general, and so perhaps present in all psychotherapeutic and contemplative practices that lead to increased well-being (to a lesser or greater degree).

A first key mechanism we consider is the case where the disruption or suspension of habitual avoidant mental actions causes unconscious contents — understood as those contents withheld from awareness through systematic and motivated inattention — to emerge into conscious awareness. This can include repressed emotions, unresolved conflicts, and concealed beliefs, all of which may have been generating (unresolveable) stress and impose an allostatic load on the cognitive system.

The disruption or suspension of these suppressive mechanisms can have several benefits. In the previous section, we argued that the suppressive effect of motivated inattention hinders updating of the generative model, causing a failure to accurately predict the consequences of action. This divergence accumulates in a growing isolation of the prior preferences. Disrupting the mechanisms of motivated inattention exposes the generative model to previously avoided sensory evidence, allowing for its updating.The previously blocked updating of the generative model will then cause an increase in affective valence. Recall, affective valence tracks the confidence or precision assigned to various beliefs. There are two ways to minimize variational free energy: bring the world into line with the generative model via action, or bring the generative model in line with the world through inference and learning. Acutely, this process may incur some pain or cognitive dissonance (i.e. decrease in affective valence), because the sensory evidence being attended to will reduce model precision. Over time, however, new confidence will emerge as the model comes into alignment with the environment. And indeed, the eventual 'insight' may feel good, just as a positive mood can promote insight (Watts et al, 2017). This release of unconscious content can be simultaneously experienced as painful and cathartic, considering both the drop and gain of model precision involved in this process.

A second key mechanism related to increasing subjective well-being is improved emotional granularity. Recall, motivated inattention acts to attenuate interoceptive evidence of stress, particularly from interoceptive channels. Suspension of habitual avoidant mental actions allows the suspended evidence to surface, allowing it to be assimilated during the interoceptive inference that underwrites emotional experience. The result is that the system is able to make more granular emotional inference — in other words, the system has a more fine grained, expressive (self) model as to the causes of allostatic load or stress, and the epistemic gain in perceived allostatic self-efficacy in having increased confidence as to actions that would resolve the stress. Furthermore, to the extent that the improved emotional granularity is reliable and informative, it allows the agent to reach preferred states more efficiently. The above scenario can be thought of as consistent with epistemic growth or development via psychotherapy, where the patient develops a more nuanced and comprehensive understanding of themselves and their unconscious. Computationally this can be understood as a process of structure learning or Bayesian model selection (Friston et al., 2023), whereby the agent increases the complexity of their model by, for instance, adding additional state levels to an emotional state factor to account for the fine-grained interoceptive evidence that is now unattenuated.

Lastly, the improved inferential construction of emotions — that results from attending to previously attenuated evidence — can lead to insights which in turn allow the agent to gain control over its generative model. For example, identifying a specific job or relationship as the source of unhappiness enables individuals to undertake targeted actions aimed at improving their life circumstances. Another example would be to notice one's own self destructive thoughts or behaviors — that can only be acted upon once they are identified. Recognition of the underlying causes of one's unhappiness can, in the short term, bring about psychological challenges. For example, if an agent realizes that they are in a destructive environment (e.g. a particularly bad relationship, family, or job) and not able or willing to change this environment, this can lead to a feeling of helplessness or cognitive dissonance. However, in the long run, these insights are likely to allow the agent to gain control by better understanding the relationships between the environmental conditions and their subjective states of well-being, culminating in tangible improvements in their psychological and emotional well-being. Given that this process can lead to a temporary feeling of helplessness if too many insights emerge too quickly, it may be beneficial to break down the process of motivated inattention slowly and gently to arrive at insights without overwhelm.

Section 8: Breaking canalization: Psychedelic therapy and meditation

The approach outlined in this paper gives a novel lens on how to consider a wide range of psychotherapeutic and contemplative practices, particularly those that involve releasing contents from the dynamic unconscious by changing patterns of attention.

There has been resurgence of interest in the therapeutic potential of psychedelic substances in recent years (Nichols, 2016). A growing body of evidence suggests that these compounds may have significant benefits for various mental health conditions, including depression (Carhart-Harris & Goodwin, 2017), anxiety (Gasser et al., 2014), post-traumatic stress disorder (PTSD) (Mithoefer et al., 2018), and substance use disorders (Bogenschutz & Johnson, 2016). The dominant theoretical framework for interpreting the psychoactive effects of psychedelics and their therapeutic potential is the REBUS (Relaxed Beliefs Under Psychedelics) model of psychedelic action (Carhart-Harris & Friston, 2019) — a theory of the mechanistic underpinnings of the mind-altering effects of psychedelic substances within the predictive brain. The REBUS model posits that psychedelics induce alterations in the hierarchically organized structure of brain activity that is thought to encode predictive models, prior beliefs or assumptions. More specifically, it is proposed that via a dysregulating effect on spontaneous cortical activity i.e., the so-called 'entropic brain effect', psychedelics reduce the precision-weighting of (particularly) high-level priors enabling a relative freeing of bottom-up information flow.nd that this process is bidirectional i.e., that liberated bottom-up information flow causes and is caused by more entropic activity in the cortex (Carhart-Harris & Friston, 2019).

As we have seen, canalization can be understood in terms of dynamic attractor states — states the system tends to revisit or converge to. This can be represented as an energy landscape whereby deep valleys or canals represent attractors. The deeper the canals, the more resistant the state is to perturbation. Within the active inference framework, the landscape is a free energy landscape and active inference represents gradient descent, where the steepness — technically, curvature — of the valleys corresponds to precision. These precise models or valleys can be understood to correspond to behavioral and cognitive phenotypes, where the higher the precision (the deeper and steeper the valley) the more resistant it is to outside (or indeed inside) perturbations and change. Psychedelics, then, are understood to flatten this landscape, enabling less hierarchically organized and regulated information flow (Carhart-Harris & Friston, 2019; Hipolito et al, 2023).

This flattening allows for a break from the usual patterns, enabling the system to explore a wider array of states. This is particularly significant when considering the therapeutic potential of psychedelics. By disrupting canalized patterns of thinking and behaving, psychedelics offer an opportunity for individuals to escape from rigid cognitive and behavioral loops, potentially leading to transformative insights and healing. In short, the mechanism of therapeutic action of psychedelic experience may, via the relaxation of avoidant and repressive mental actions, allow unconscious contents to be brought into awareness, which, if well processed e.g., with psychotherapeutic support, may drive therapeutic outcomes via an epistemic development.

Evidence has increasingly linked meditation with sustained improvements in subjective well-being. Meditation practices are rooted in ancient wisdom traditions, and involve focusing the mind and cultivating awareness in order to achieve mental clarity, emotional stability, and what has been described in some traditions as “spiritual enlightenment” (Lutz, Slagter, Dunne & Davidson, 2008; Sparby & Sacchet, 2022). Meditation is thought to provide numerous benefits, including improved concentration, reduced stress, and enhanced emotional well-being (Tang, Hölzel & Posner, 2015). A growing body of scientific research supports these claims, showing that meditation can foster positive changes in brain structure and function, leading to increased self-awareness, emotional regulation, and overall improved subjective well-being (Sezer et al, 2022). More recently there is growing interest in further explicating advanced meditation including the developmental trajectory toward limits and endpoints of practice (Galante et al, 2023; 2023; Wright et al, 2023; Sparby & Sacchet, 2022; Yang et al, 2023). It worth noting that there is evidence suggesting meditation is not always associated with improved subjective well-being (Britton, 2019), as there is with psychedelic use or therapy (Schlag et al, 2022),

Active inference has been increasingly applied to the study of meditation (Pagnoni & Guareschi, 2017, Deane et al, 2020; Laukkonen & Slagter, 2021; Laukkonen et al, 2023; Prest, 2021; Sandved-Smith et al, 2022). Lutz et al. (2019) offered the first account of focused attention meditation based on active inference principles, identifying two interrelated aims: regulating attention on a specific object (concentration or ‘Shamata’ Meditation) and enhancing one's understanding of the meditative object and various distractions, especially recognizing their dynamic and impersonal nature (Insight or Vipassana; meditation).

According to active inference, focused attention meditation requires top-down precision enhancement of behavioral policies associated with stable attention on an object (Lutz et al., 2019). Meditators must maintain this policy while facing competition from other simultaneously active policies (Pezzulo & Cisek, 2016). During meditation, this competition narrows down to preserving attention on the meditative object and resisting policies that divert attention, such as spontaneous memories, future planning, homeostatic concerns, or mind wandering. ‘Inaction’ is crucial in this process, as it allows the dialectic between focused attention and distraction to unfold and be consciously attended to. As meditators become more experienced, they are able to allow distracting thoughts and sensations to arise and pass without disrupting their concentration (Dambrun & Ricard, 2011; Dambrun, 2016). It has been suggested that this is partially achieved through reduced precision on the prior preferences that drive the distraction, such that they cease to draw attention in the same manner (Deane et al, 2020).

In the context of sustained attentional control, the goal is to be aware of where one's attention is focused, quickly recognize shifts in attention (i.e., distractions), and recalibrate attentional processes accordingly back to the meditation object (such as the breath). During focused attention and mindfulness meditation practices, it’s typical to have phenomenological cycles of focus and mind-wandering long associated with (Lutz et al. 2008). Sandved-Smith et al., (2021) utilize a model of meta-awareness and attentional control based on hierarchical active inference to formalize this process. This model views mental action not only as policy selection over higher-level cognitive states, as described earlier in this paper, but also includes an additional hierarchical level representing meta-awareness states. These meta-awareness states regulate the expected confidence (precision) in the relationship between observations and underlying cognitive states.

The key mechanism underlying the increment in subjective well-being associated with meditation practice may be the suspension of the habitual avoidant mental actions, which, if successful, instigates a release of ‘unconscious contents’. In the absence of potential overt actions (moving around, scrolling social media, etc), covert actions such as mind-wandering may take place as a means to attenuate the inputs associated with negative affect (such as the specific memory, interoceptive sensations, etc). As discussed above, this usually serves the function of maintaining an affective state. Meditation, in simply drawing attention back to the meditation object, disrupts these habitual processes, such that the evasive or repressive mental actions are suspended. As such, content that had previously been selectively disattended to, can rise up and register within perceptual inference and awareness. Furthermore, by cultivating the capacity for meta-awareness of the previously transparent processes driving the inattention, the practitioner's generative model is capable of inferring the suboptimal ramifications of canalized inattention and can therefore pursue a path informed by greater self understanding. While prior studies, including related to advanced meditative states that are characterized by the radical breakdown of priors and consciousness more generally, have provided neuroscientific evidence that can be interpreted meaningfully within this context (Yang et al., 2023; Chowdhury et al., 2023), future research could more directly test this account of meditation.

Conclusion

This paper has introduced the notion of adaptive narrative control and articulated it within the active inference framework, with a focus on how particular kinds of adaptive narrative control can lead to psychopathology. This furnishes a novel perspective on the so-called canalization model of psychopathology, a theory on how a putative principal component of psychopathology or ‘p-factor’ may be generated. Our account chimes with the Freudian view defense mechanisms being at the heart of psychopathology, as well as more contemporary work on psychological avoidance (Hayes et al, 2006).

This is a preliminary picture, and has not applied the theoretical framework of adaptive narrative control to understanding the etiology of psychopathology. Future work needs to identify the dynamics and developmental trajectories underwriting particular psychopathologies, as understood as driven by adaptive narrative control. For instance, an account of delusion formation on this picture would suggest that an imperative to regulate affective states leads to increasingly unusual sampling of available evidence through attentional control. This is likely due to particular environmental conditions which necessitate affective regulation via particular kinds of narrative control.

It is also likely that a system engages in compensatory strategies such as narrative control in cases where it is unable to ‘unlearn’ a relationship between a stimulus and response quickly enough in order to function in a given context. For instance, we could consider a case of a child who has learned crying is an effective way to get their needs met. If their environment changes to one where crying is met with social punishment, it may be easier (i.e. faster and computationally cheaper) for the system to attenuate evidence leading to the crying (e.g. attenuate interoceptive inputs signaling hunger) than it is to unlearn that crying is a viable way of getting needs met. Adaptive narrative control, then, comes at the expense of not unlearning particular patterns of responses to the environment, as withdrawing attention prevents the system from encountering the disconfirmatory evidence required to revise the original belief structures. As such, layers of compensatory attentional mechanisms (‘kludges’) may accrue over the course of development. The dynamics of adaptive narrative control is an important question for future work.

Future work should also specify therapeutic mechanisms with reference to particular disorders, in order to refine a picture of appropriate therapeutic intervention. We have not touched on how talking therapy undo the pathological side of canalisation. One simple idea would be that it allows remembering and enacting of the same previously repressed content that led to the mental avoidance in the first place, but now in a different (safer) context facilitating the co-creation of new narratives with more granular awareness of interoceptive states and viable behavioral responses.

There are also broader questions that have not been touched upon: if psychopathology and self-deception come as a consequence of systems becoming more sophisticated, hierarchical, and reflective; which intelligent systems (either biological or artificial) are vulnerable to such processes, and under what conditions? Could a synthetic intelligence be prone to avoidant defense mechanisms as humans seem to be, and could this be prevented, without compromising its sophistication or intelligence? In other work (in preparation) we make the case that any system with an attention schema and a value function is susceptible to adaptive narrative control, highlighting the possibility of self-deceptive artificial intelligence.

Finally, there are numerous philosophical implications that remain unexplored. Our account introduces a mechanism as to the gap between an agent's narrative about the causes of their sensations, and the ‘true’ or underlying causes, where false narratives can confer an adaptive advantage in certain contexts. A potential conception of free will, in the vein of philosopher Harry Frankfurt, is to understand an agent as ‘freer’ the smaller this gap is between their narrative about the factors driving their behavior, and what is ‘really’ driving behavior. We leave these questions to future work.

References

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical psychology review*, *30*(2), 217-237.

Arnaldo, I., Corcoran, A. W., Friston, K. J., & Ramstead, M. J. (2022). Stress and its sequelae: An active inference account of the etiological pathway from allostatic overload to depression. *Neuroscience & Biobehavioral Reviews*, *135*, 104590.

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of psychosomatic research*, *38*(1), 23-32.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1-23.

Belsky, J., & Pluess, M. (2013). Beyond risk, resilience, and dysregulation: Phenotypic plasticity and human development. *Development and psychopathology*, *25*(4pt2), 1243-1261.

Blakemore, S. J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself?. *Neuroreport*, *11*(11), R11-R16.

Bogenschutz, M. P., & Johnson, M. W. (2016). Classic hallucinogens in the treatment of addictions. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 64, 250-258.

Britton, W. B. (2019). Can mindfulness be too much of a good thing? The value of a middle way. *Current opinion in psychology*, *28*, 159-165.

Carhart-Harris, R. L., & Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain*, *133*(4), 1265-1283.

Carhart-Harris, R. L., & Goodwin, G. M. (2017). The therapeutic potential of psychedelic drugs: Past, present, and future. Neuropsychopharmacology, 42(11), 2105-2113.

Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, *71*(3), 316-344.

Carhart-Harris, R. L., Chandaria, S., Erritzoe, D. E., Gazzaley, A., Girn, M., Kettner, H., ... & Friston, K. J. (2022). Canalization and plasticity in psychopathology. *Neuropharmacology*, 109398.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... & Moffitt, T. E. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders?. *Clinical psychological science*, *2*(2), 119-137.

Chowdhury, A., van Lutterveld, R., Laukkonen, R. E., Slagter, H. A., Ingram, D. M., & Sacchet, M. D. (2023). Investigation of advanced mindfulness meditation "cessation" experiences using EEG spectral analysis in an intensively sampled case study. *Neuropsychologia*, *190*, 108694.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

Cicchetti, D., & Tucker, D. (1994). Development and self-regulatory structures of the mind. *Development and psychopathology*, *6*(4), 533-549.

Constant, A. (2021). The free energy principle: it's not about what it takes, it's about what took you there. Biology & Philosophy, 36(2), 10.

Da Costa, L., Friston, K., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A*, *477*(2256), 20210518.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 473-492.

DeWitt, T. J., Sih, A., & Wilson, D. S. (1998). Costs and limits of phenotypic plasticity. *Trends in ecology & evolution*, *13*(2), 77-81.

Dambrun, M., & Ricard, M. (2011). Self-centeredness and selflessness: A theory of self-based psychological functioning and its consequences for happiness. *Review of General Psychology*, *15*(2), 138-157.

Dambrun, M. (2016). When the dissolution of perceived body boundaries elicits happiness: The effect of selflessness induced by a body scan meditation. *Consciousness and cognition*, *46*, 89-98.

Deane, G. (2020). Dissolving the self: Active inference, psychedelics, and ego-dissolution. *Philosophy and the Mind Sciences*, *1*(I), 1-27.

Deane, G., Miller, M., & Wilkinson, S. (2020). Losing ourselves: active inference, depersonalization, and meditation. *Frontiers in psychology*, *11*, 539726.

Deane, G. (2021). Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution. *Neuroscience of Consciousness*, *2021*(2), niab024.

Deane, G. (2022). Machines That Feel and Think: The Role of Affective Feelings and Mental Action in (Artificial) General Intelligence. *Artificial Life*, *28*(3), 289-309.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature neuroscience*, *8*(11), 1481-1489.

Flatt, T. (2005). The evolutionary genetics of canalization. *The Quarterly review of biology*, *80*(3), 287-316.

Freud, S. (1912). A note of the unconscious in psycho-analysis.

Freud, S. (1914). On narcissism.

Freud, S. (1922). The unconscious. *The Journal of Nervous and Mental Disease*, *56*(3), 291-294.

Freud, S. (2012). *The basic writings of Sigmund Freud*. Modern library.

Fotopoulou, A., Pfaff, D., & Conway, M. A. (Eds.). (2012). *From the couch to the lab: Trends in psychodynamic neuroscience*. Oxford University Press.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, *6*(4), 187-214.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, *29*(1), 1-49.

Friston, K. J., Da Costa, L., Tschantz, A., Kiefer, A., Salvatori, T., Neacsu, V., Koudahl, M., Heins, C., Sajid, N., Markovic, D., Parr, T., Verbelen, T., & Buckley, C. L. (2023). Supervised structure learning (arXiv:2311.10300). arXiv. http://arxiv.org/abs/2311.10300

Friston, K. (2018). Am I self-conscious?(Or does self-organization entail self-consciousness?). *Frontiers in psychology*, *9*, 579.

Galante, J., Friedrich, C., Dalgleish, T., Jones, P. B., & White, I. R. (2023). Systematic review and individual participant data meta-analysis of randomized controlled trials assessing mindfulness-based programs for mental health promotion. *Nature mental health*, *1*(7), 462-476.

Gasser, P., Kirchner, K., & Passie, T. (2014). LSD-assisted psychotherapy for anxiety associated with a life-threatening disease: A qualitative study of acute and sustained subjective effects. Journal of Psychopharmacology, 29(1), 57-68.

Gómez-Emilsson, A. (2021). Healing trauma with neural annealing. *QRI*.

Goodkind, M., Eickhoff, S. B., Oathes, D. J., Jiang, Y., Chang, A., Jones-Hagata, L. B., ... & Etkin, A. (2015). Identification of a common neurobiological substrate for mental illness. *JAMA psychiatry*, *72*(4), 305-315.

Guidi, J., Lucente, M., Sonino, N., & Fava, G. A. (2020). Allostatic load and its impact on health: a systematic review. *Psychotherapy and psychosomatics*, *90*(1), 11-27.

Harris, D. J., Wilkinson, S., & Ellmers, T. J. (2023). From fear of falling to choking under pressure: A predictive processing perspective of disrupted motor control under anxiety. Neuroscience & Biobehavioral Reviews, 148, 105115. https://doi.org/10.1016/j.neubiorev.2023.105115

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour research and therapy*, *44*(1), 1-25.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural computation*, *33*(2), 398-446.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Hohwy, J. (2016). The self‐evidencing brain. *Noûs*, *50*(2), 259-285.

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II).

Helmholtz, H. (1867). *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln* (Vol. 9). Voss.

Hipólito, I., Mago, J., Rosas, F. E., & Carhart-Harris, R. (2023). Pattern breaking: a complex systems approach to psychedelic medicine. *Neuroscience of Consciousness*, *2023*(1), niad017.

Ji, X., Elmoznino, E., Deane, G., Constant, A., Dumas, G., Lajoie, G., ... & Bengio, Y. (2023). Sources of richness and ineffability for phenomenally conscious states. *arXiv preprint arXiv:2302.06403*.

Kang, S. M., & Shaver, P. R. (2004). Individual differences in emotional complexity: Their psychological implications. *Journal of personality*, *72*(4), 687-726.

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological cybernetics*, *112*(4), 323-343.

Kernberg, O. (1967). Borderline personality organization. *Journal of the American psychoanalytic Association*, *15*(3), 641-685.

Kersten, L. A Model Solution: On the Compatibility of Predictive Processing and Embodied Cognition. Minds & Machines 33, 113–134 (2023). https://doi.org/10.1007/s11023-022-09617-7

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS computational biology*, *4*(11), e1000209.

Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, *220*(4598), 671-680.

Koster, E. H., De Lissnyder, E., Derakshan, N., & De Raedt, R. (2011). Understanding depressive rumination from a cognitive science perspective: The impaired disengagement hypothesis. *Clinical psychology review*, *31*(1), 138-145.

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood?. *Journal of abnormal psychology*, *121*(4), 971.

Leichsenring, F., & Rabung, S. (2011). Long-term psychodynamic psychotherapy in complex mental disorders: update of a meta-analysis. *The British Journal of Psychiatry*, *199*(1), 15-22.

Lane, R. D., & Smith, R. (2021). Levels of emotional awareness: theory and measurement of a socio-emotional skill. *Journal of Intelligence*, *9*(3), 42.

Laplanche, J., & Pontalis, J. B. (2018). *The language of psychoanalysis*. Routledge.

Lewis, M. D., & Todd, R. M. (2007). The self-regulating brain: Cortical-subcortical feedback and the development of intelligent action. *Cognitive Development*, *22*(4), 406-430.

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in cognitive sciences*, *12*(4), 163-169.

Lutz, A., Mattout, J., & Pagnoni, G. (2019). The epistemic and pragmatic value of non-action: a predictive coding perspective on meditation. *Current opinion in psychology*, *28*, 166-171.

Mann, S. F., Pain, R., & Kirchhoff, M. D. (2022). Free energy: a user's guide. *Biology & Philosophy*, *37*(4), 33.

Maroti, D., Lilliengren, P., & Bileviciute-Ljungar, I. (2018). The relationship between alexithymia and emotional awareness: a meta-analytic review of the correlation between TAS-20 and LEAS. *Frontiers in psychology*, *9*, 453.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. mit Press.

Metzinger, T. (2017). "The problem of mental action: predictive control without sensory sheets," in Philosophy and Predictive Processing, eds T. Metzinger and W. Wiese (Frankfurt am Main: MIND Group), 1–26.

Miller, W. R., & C'de Baca, J. (2001). *Quantum change: When epiphanies and sudden insights transform ordinary lives*. Guilford Press.

Mithoefer, M. C., Mithoefer, A. T., Feduccia, A. A., Jerome, L., Wagner, M., Wymer, J., ... & Doblin, R. (2018). 3,4-methylenedioxymethamphetamine (MDMA)-assisted psychotherapy for post-traumatic stress disorder in military veterans, firefighters, and police officers: A randomised, double-blind, dose-response, phase 2 clinical trial. The Lancet Psychiatry, 5(6), 486-497.

Nave, K., Deane, G., Miller, M., & Clark, A. (2022). Expecting some action: Predictive Processing and the construction of conscious experience. *Review of Philosophy and Psychology*, *13*(4), 1019-1037.

Nichols, D. E. (2016). Psychedelics. *Pharmacological reviews*, *68*(2), 264-355.

Pagnoni, G., & Guareschi, F. T. (2017). Remembrance of things to come: a conversation between Zen and neuroscience on the predictive nature of the mind. *Mindfulness*, *8*, 27-37.

Palmer, C. E., Davare, M., & Kilner, J. M. (2016). Physiological and perceptual sensory attenuation have different underlying neurophysiological correlates. *Journal of neuroscience*, *36*(42), 10803-10812.

Patel, P., Hancock, J., Rogers, M., & Pollard, S. R. (2022). Improving uncertainty tolerance in medical students: A scoping review. *Medical Education*, *56*(12), 1163-1173.

Pauls, D. L., Abramovitch, A., Rauch, S. L., & Geller, D. A. (2014). Obsessive–compulsive disorder: an integrative genetic and neurobiological perspective. *Nature Reviews Neuroscience*, *15*(6), 410-424.

Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, *20*(6), 414-424.

Prest, T. S. G. Opening the Sensory Gates.

Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of experimental psychology: General*, *143*(6), 2057.

Ramstead, M. J., Wiese, W., Miller, M., & Friston, K. J. (2020). Deep neurophenomenology: An active inference account of some features of conscious experience and of their disturbance in major depressive disorder.

Ramstead, M. J., Sakthivadivel, D. A., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., ... & Friston, K. J. (2023). On Bayesian mechanics: a physics of and by beliefs. *Interface Focus*, *13*(3), 20220029.

Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K., Lutz, A., & Ramstead, M. J. (2021). Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of consciousness*, *2021*(1), niab018.

Schlag, A. K., Aday, J., Salam, I., Neill, J. C., & Nutt, D. J. (2022). Adverse effects of psychedelics: From anecdotes and misinformation to systematic science. *Journal of Psychopharmacology*, *36*(3), 258-272.

Sezer, I., Pizzagalli, D. A., & Sacchet, M. D. (2022). Resting-state fMRI functional connectivity and mindfulness in clinical and non-clinical contexts: A review and synthesis. *Neuroscience & Biobehavioral Reviews*, *135*, 104583.

Sharot, T. (2011). The optimism bias. Current biology, 21(23), R941-R945.

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., ... & Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological bulletin*, *144*(4), 343.
Singleton, S. P., Luppi, A. I., Carhart-Harris, R. L., Cruzat, J., Roseman, L., Nutt, D. J., ... & Kuceyeski, A. (2021). LSD and psilocybin flatten the brain's energy landscape: insights from receptor-informed network control theory. BioRxiv, 2021-05.

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of mathematical psychology*, *107*, 102632.

Smith, R., & Lane, R. D. (2015). The neural basis of one's own conscious and unconscious emotional states. *Neuroscience & Biobehavioral Reviews*, *57*, 1-29.

Smith, R., & Lane, R. D. (2016). Unconscious emotion: A cognitive neuroscientific perspective. *Neuroscience & Biobehavioral Reviews*, *69*, 216-238.

Smith, R., Killgore, W. D., & Lane, R. D. (2018). The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*, *18*(5), 670.

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Frontiers in psychology*, *10*, 2844.

Smith, R., Varshney, L. R., Nagayama, S., Kazama, M., Kitagawa, T., & Ishikawa, Y. (2022). A computational neuroscience perspective on subjective wellbeing within the active inference framework. *International Journal of Wellbeing*, *12*(4).

Solms, M. (2017). What is "the unconscious," and where is it located in the brain? A neuropsychoanalytic perspective. *Annals of the New York Academy of Sciences*, *1406*(1), 90-97.

Sparby, T., & Sacchet, M. D. (2022). Defining meditation: foundations for an activity-based phenomenological classification system. *Frontiers in psychology*, *12*, 795077.

Speechley, W. J., Whitman, J. C., & Woodward, T. S. (2010). The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience*, *35*(1), 7-17.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., ... & Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in human neuroscience*, *10*, 550.

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiology & behavior*, *106*(1), 5-15.

Sydnor, V. J., Larsen, B., Bassett, D. S., Alexander-Bloch, A., Fair, D. A., Liston, C., ... & Satterthwaite, T. D. (2021). Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology. *Neuron*, *109*(18), 2820-2846.

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

Trevisan, D. A., Altschuler, M. R., Bagdasarov, A., Carlos, C., Duan, S., Hamo, E., ... & McPartland, J. C. (2019). A meta-analysis on the relationship between interoceptive awareness and alexithymia: Distinguishing interoceptive accuracy and sensibility. *Journal of Abnormal Psychology*, *128*(8), 765.

Velasco, P. F. (2017). Attention in the Predictive Processing Framework and the Phenomenology of Zen Meditation. *Journal of Consciousness Studies*, *24*(11-12), 71-93.

Vocks, S., Busch, M., Schulte, D., Grönermeyer, D., Herpertz, S., & Suchan, B. (2010). Effects of body image therapy on the activation of the extrastriate body area in anorexia nervosa: an fMRI study. *Psychiatry Research: Neuroimaging*, *183*(2), 114-118.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, *150*(3811), 563-565.

Waddington, C.H., 1959. Canalization of development and genetic assimilation of acquired characters. Nature 183 (4676), 1654–1655.

Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., & Barrett, L. F. (2015). A Bayesian model of category-specific emotional brain responses. *PLoS computational biology*, *11*(4), e1004066.

Watts, R., Day, C., Krzanowski, J., Nutt, D., & Carhart-Harris, R. (2017). Patients' accounts of increased "connectedness" and "acceptance" after psilocybin for treatment-resistant depression. Journal of humanistic psychology, 57(5), 520-564.

Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2023). A critical evaluation of the p-factor literature. *Nature Reviews Psychology*, 1-15.

Wiese, W. (2017). Action is enabled by systematic misrepresentations. *Erkenntnis*, *82*(6), 1233-1252.

Wright, M. J., Sanguinetti, J. L., Young, S., & Sacchet, M. D. (2023). Uniting Contemplative Theory and Scientific Investigation: Toward a Comprehensive Model of the Mind. *Mindfulness*, 1-14.

Yang, W. F. Z., Chowdhury, A., Bianciardi, M., van Lutterveld, R., Sparby, T., & Sacchet, M. D. (2023). Intensive whole-brain 7T MRI case study of volitional control of brain activity in deep absorptive meditation states. *Cerebral Cortex*, bhad408.