# The Shape of Subjectivity:

## An active inference approach to consciousness and altered self-experience

George Deane

PhD in Philosophy

The University of Edinburgh

2021

# Declaration of Authorship

I, George Deane, declare that this thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below (described in "Author's Contributions"). I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

George Deane, February 19th 2021

# Acknowledgements

I particularly want to thank my primary supervisor, Andy Clark, for guidance in writing this thesis and for such illuminating and fun conversations throughout.

This PhD has been part of the ERC funded XSPECT project. Many thanks to all of the XSPECT team, past and present: Sam Wilkinson, Mark Miller, Kate Nave, David Carmel, Frank Schumann and Andy Clark.

Thank you to my secondary supervisor, Mark Sprevak, for being incredibly helpful and supportive, and for taking me on at a late stage in my PhD.

Many thanks to the friends and colleagues at the University of Edinburgh and beyond for the interesting discussions that have shaped this thesis, and for making this such a happy period of my life: David Harrison, Fausto Carcassi, Giles Howdle, Matt Sims, Urtė Laukaitytė, Luke Kersten, Roberta Leotta, Suraiya Lueke, Lina Skora, Stavros Anagnou, Jenny Zhang, James Brown, Sophie Potter, Liv Coombes, Camden McKenna, Julia Lubner, Tijana Jokic, Viky Neascu, Giada Margiotto, Nina Poth, Julian Hauser, Danaja Rutar, Kate Nave, Kate Webb, Jan Meyer, Ana Pujic, Keisuke Suzuki, Kris Moody, Jonas Mago, Jonas Nölle, Mel Andrews, Nicole Chan, Neil Bramley, Shannon Proksch, Matthew Crosby, Becky Millar, Bruno Savill De Jong, Marise Treseder and Lucy Allan.

Thanks to my parents for their ongoing support and encouragement. Thanks to my sister Elisabeth and brother-in-law Jethro, particularly for their kindness in inviting me to stay with them. Much love to my Auntie Sue and Granny-Peg, who we sadly lost during this period.

I am very grateful to have the opportunity to engage what has been a primary passion—seeking to understand the nature of the mind and consciousness. I dedicate this thesis to Mimi, my grandma who passed away in February of 2020. Her kindness and support throughout my life have been invaluable to me, and her sense of fun and curiosity remain a great inspiration.

# Abstract/Lay summary

How should we understand the place of the mind in the natural world? Can the relationship between the contents of consciousness and the underlying mechanisms be identified? This thesis approaches the question of consciousness and the self through the framework of active inference. According to predictive processing approaches to brain function, brains are essentially prediction machines. On this view, perception and action are underpinned by inferential mechanisms that implement a hierarchical generative model, constantly attempting to match incoming sensory inputs with top-down predictions or expectations. Predictive processing is thought to offer a first glimpse of a unified theory of the mind—uniting perception, action, and cognition under a single theoretical framework. In particular, active inference, under the free energy principle, has emerged as the most explanatorily powerful approach in predictive processing. In this thesis, I develop a conceptual framework within active inference for understanding consciousness and phenomenal selfhood (broadly, the 'sense of being a self') in terms of an "allostatic control model". I made the case that phenomenal selfhood arises from a hierarchically deep inference about endogenous control of 'self-evidencing' (survival-relevant) sensory outcomes.

I apply this account to develop a new understanding of the relationship between self-consciousness and consciousness. Based on the allostatic control model, I posit a novel theoretical model of how psychedelic drugs can lead to 'selfless' experiences. I then apply the allostatic control model to characterise the contrastingly dysphoric and euphoric selfless experiences that can arise in depersonalisation disorder and meditation practice. Based on these accounts, I consider the possibility of a theory of consciousness within this active inference, analysing whether selfless experiences pose a threat to an active inference theory of consciousness understood in terms of self-modelling mechanisms. I argue that selfless experiences do not pose a threat to an active inference theory of consciousness, rather selfless states can be informative as to how consciousness should be understood in active inference. Consciousness emerges as fundamentally affective on this view, where (in normal experience) hierarchically deep self-modelling mechanisms function to 'tune' organisms to opportunities for adaptive action across multiple interlocking timescales.

# Table of Contents

# Introduction

## The Astonishing Hypothesis

Francis Crick's 'Astonishing Hypothesis', that: "You, your joys and sorrows, your memories and ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules" (1994, p3), is a great mystery of science and philosophy. How is it that certain arrangements of matter—such as the electrochemical activity in brain—give rise to experience? The approach of this thesis is to locate conscious experience in an existing overarching theory of brain function. In recent years 'predictive processing' theories are increasingly taking precedence as the guiding theoretical models in brain science.

## Predictive processing

Predictive Processing (PP) casts the brain as a probabilistic prediction machine that continually generates top-down predictions of the hidden causes of the impinging sensory inputs. Predictive processing theories such as the free-energy principle, and its process theory active inference, are thought to herald a first glimpse of a unified theory of cognition, action and perception (Clark, 2013; Friston, 2010). On this view, the brain implements a hierarchical generative model, aimed at recapitulating the causal structure of the world, in order to infer the hidden causes of sensation. This model is refined over the course of phylogeny and ontogeny, iteratively sculpted through incoming prediction error (mismatches between the predictions and the incoming signal) flowing from the sensory peripheries. The generative model is hierarchically organised: lower sensory areas are provided with predictions (or 'priors' in Bayesian terms) from higher cortical regions. The generative model is then updated in light of the prediction error—the more surpising the sensory input the more revision is required. PP inverts the standard picture of perception as a kind of 'imprinting' on a passive brain, and instead construes it as a kind of controlled hallucination—where top-down predictions are supervised by incoming sensory evidence.

Perceptual inference results from the system seeking to minimise prediction error across the hierarchy so the levels are in agreement. This means the system is constantly revising and updating itself in light of the incoming sensory flow. There are clues here as to where consciousness fits into the PP picture – consciousness emerges as the “upshot” of unconscious perceptual inference – the hypothesis that best minimises prediction error spanning multiple hierarchichal levels:

> “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. More specifically, since the inversion of the generative model is implicit, what is conscious is the interconnected set of currently best performing predictions down throughout the perceptual hierarchy (down to some level of expected fineness of perceptual grain)” (Hohwy, 2013, p 201)

Similarly, Clark (2012) states:

> "Experience is conditioned upon the best linked set of hypotheses spanning multiple spatial and temporal scales (given current context and accommodating the driving sensory signal)" (P13).

The predictive processing framework has been applied to account for an impressive range of phenomenological states. For instance, in object recognition, the 'gist' of the scene engages past experience (activates an associative network of priors) to generate the most likely prediction about the object's identity (Bar, 2003; Oliva & Torralba, 2007). For instance, in the case of ambiguous input the context of a scene can determine whether an object is perceived as a hairdryer or a drill depending on whether the context is in a bathroom or a workshop. These predictions are fed back to early visual areas to speed perception by constraining the hypothesis space. This scene perception can be understood as an instance of a more general feature of the predictive brain—that experience activates 'associative networks' that constrain the hypothesis space in order to make better predictions and minimise prediction error (Aminoff & Tarr, 2015; Bar, 2004; Summerfield & Egner, 2009). Priming effects can be understood as another manifestation of this same mechanism.

Can predictive processing provide us with a theory of consciousness? Consciousness within PP is typically not considered in terms of the construction of a subjective perspective. What is missing from the current picture is a view of why there is an experiencing subject, why there is *something-it-is-like* for that system to perceive. Supplementing this picture with two primary sources—embodiment and action—can make strides in this direction. Approaches in 'embodied cognition' unify the ways the body, action, and the use of environmental structure are used to simply the cognitive challenges faced by the brain (Pfeifer & Bongard, 2006; Clark, 1997; Brooks, 1991). The last two decades has seen increasing emphasis on embodied cognition, highlighting that perception is deeply 'action-oriented'. (Engel et al, 2013). Perception and neural representations are geared towards adaptive actions, engaging the organisms in action opportunities conducive to survival and reproduction. Limited processing capacity means that these representations will be as minimal as possible, only as rich as is necessary in order to successfully complete the action, such that more costly information processing is not engaged when simpler routines are able to do the job. Increasingly, the importance of embodiment is being emphasised in the PP literature (Seth & Tsakiris, 2018), particularly within the free energy principle under active inference (e.g. Pezzulo et al., 2015)

## The free energy principle

In recent years, more embodied and action-oriented approaches have taken precedence in the predictive processing literature, in the *active inference* framework. At the centre of the active inference framework is the free energy principle (FEP), an ambitious unifying principle that combines and subsumes numerous approaches to the brain, including the Bayesian brain (Knill & Pouget, 2004), predictive coding (Mumford, 1992; Rao & Ballard, 1999) reinforcement learning (Dayan & Daw, 2008) and efficient coding (Barlow, 2001). At the heart of the free energy principle lays the autopoietic principle that self-organising systems resist the natural tendency to disorder implied by the second law of thermodynamics, by keeping their internal states in a state of equilibrium in the face of an ever-changing environment.

Minimising free energy can be understood as minimising the evidence of one's own dispersion, formally equivalent to maximising Bayesian model evidence, and as such is also called 'self-evidencing' (Hohwy, 2016). That organisms regulate their states to stay viable is intuitive – animals and people act in order to stay within their "species-specific window of viability" (Clark, 2013, p.

13). According to the free energy principle, organisms do this by acting to stay in 'predictable' states, where what is predictable to an organism is phenotype specific, and most critically related to maintaining homeostatic setpoints—for example, for most land animals that would include staying in a well oxygenated environment. More formally, this imperative to resist entropy and stay viable is achieved by minimising the long-term average of information theoretic surprise. Importantly, this is beyond epistemic access to an organism, and so organisms minimise an upper bound or proxy variable – the (variational) free energy. While the mathematics of the theory is complex (Buckley, Kim, McGregor, & Seth, 2017), the core idea is simple: organisms come phylogenetically equipped with expectations for their continued existence, and then act to make these expectations come true.

The fundamental imperative of the free energy principle, then, is to minimise surprise (Friston, 2010). Continued existence thus necessitates that an organism maintains itself within a limited repertoire of phenotype-congruent states, echoing earlier control theoretic principles of cybernetics (Wiener, 1948). On this view, a system can resist perturbation to internal states—those tracking 'essential variables' (Ashby, 2013) by acting to restore the expected sensory input, where an internal reference point (also known as a setpoint or goal signal) is compared to the current state, and the system acts so as to restore conditions to the setpoint.

This connects closely to a control theoretic precursor of active inference, perceptual control theory (Powers & Powers, 1973), which casts the action as the 'control of perception'. In Powers' (1973) words: "The reference signal is a model inside the behaving system against which the sensor signal is compared: behaviour is always such as to keep the sensor signal close to the setting of this reference signal." Contemporary active inference formulations operate with very similar principles. For instance, in control-oriented predictive regulation (instrumental active inference) (Seth & Tsakiris, 2018), the setpoints thought as tracking key homeostatic variables are accessed inferentially, as internal variables are tracked via 'interoception'—the sense of the body 'from within' (Craig, 2003).

The system must then infer actions that bring these variables into reasonable bounds, keeping the organism viable and alive. Here, rather than simply inferring states of the world, the organism is act to make certain states true, such as the control of bodily states (Seth, 2014; Seth & Friston, 2016). Active inference can thus be understood as following control theoretic principles, where the essential variables are high precision prior preferences ('goal priors') of states the organism expects to be in, and seeks to realise through action. These expectations are largely unamenable to revision and phylogenetically endowed (although over the course of ontogeny the organism acquires higher-level goal priors that are predictive of homeostatic outcomes). In active inference, rather than fixed set points, prior expectations about essential variables encode probability distributions over states. This means that the sufficient statistics specifying the set point (mean and precision) can be toggled contextually and are free to vary (Ainley, Apps, Fotopoulou, & Tsakiris, 2016). Active inference formalises homeostatic control in control theoretic terms, where homeostasis can be restored not only by, for instance, autonomic reflexes such as sweating to cool down, but also by prospective control that anticipates departures from homeostatic bounds before they arise and acts to avoid them. This occurs both through interoceptive inference on current bodily states, and inference on the expected future evolution of bodily states contingent on certain actions (Pezzulo et al., 2015; A. Seth, 2014; Sterling, 2012). This process of anticipatory action, by which the brain regulates the needs of the body, is known as allostasis (Corcoran, Pezzulo, & Hohwy, 2020) Active inference formally articulates allostasis, such that agents anticipate surprising outcomes before they arise, and act in order to minimise uncertainty about potential future outcomes (Pezzulo et al., 2015; Pezzulo, Rigoli, & Friston, 2018; Schulkin & Sterling, 2019).

## Active inference

In active inference, action selection and action planning is cast as a problem of inference, whereby the system needs to select actions which minimise the 'expected free energy' associated with a given action sequence (a 'policy'). The expected free energy is the average free energy the system expects to accrue in pursuing a particular policy. Action understood as an inference problem is known as *planning as inference* (Attias, 2003; Botvinick & Toussaint, 2012; Kaplan & Friston, 2018). Selecting policies that minimise the average free energy over the long term, and therefore maximise the probability of existing, requires the balancing the pragmatic and epistemic affordance of action. The pragmatic or instrumental value is simply the probability of resulting in expected sensory states, under some prior preferences (e.g. being satiated, having a comfortable body temperature). Epistemic value refers to the expected information gain (equivalent to the reduction in uncertainty) afforded by an action policy (Kaplan & Friston, 2018). Crucially, epistemic action of this kind increases the agent's 'grip' on the environment (Bruineberg & Rietveld, 2014; Kiverstein, Rietveld, & Miller, 2017). This formulation accounts for the information-seeking behaviour of agents and explains behaviour corresponding to novelty seeking and curiosity (Friston et al., 2015; Kaplan &

Friston, 2018; Mirza, Adams, Mathys, & Friston, 2016). On this picture, the resolution of uncertainty underpins intrinsic motivation, where epistemic gain is independent of expected utility (Barto, Mirolli, & Baldassarre, 2013; Friston et al., 2015). On this view, many if not most actions are driven by a motivation to resolve uncertainty about the consequences of our actions (Oudeyer, Kaplan, & Hafner, 2007; Schmidhuber, 2010). Active inference formulations of planning an navigation have be used to dissolve the 'explore-exploit' dilemma, as the agent simply needs to act so as to minimise uncertainty (i.e. expected surprise or free energy) (Kaplan & Friston, 2018).

The question arises, then, what the active inference framework has to tell us about what it means to be an experiencing subject, over and above the predictive processing framework already described. The central argument here is that computational self-modelling mechanisms that are inherent in the active inference framework should be identified with the phenomenology of being a self. On this view, hierarchically deep contextualisation of interoceptive signals "shapes" subjectivity. Intuitively, inference about divergence from certain expected set points manifests experientially as hunger, and this can modulate the salience of perceptual inputs—such as the smell of freshly baked bread.

Along these lines, Montague and King-Casas (2007) write:

> "A sated and comfortable lioness looking at two antelopes sees two unthreatening creatures against the normal backdrop of the temperate savannah....The same lioness, when hungry, sees only one thing—the most immediate prey. . . . In another circumstance, in which the lioness may be inordinately hot, the distant, shaded tree becomes the prominent visual object in the field of view. (p. 519)

This simple example illustrates how inference about the state of the body permeates perception, and perceptual salience is determined accordingly:

> "[T]he mismatch between the internal need (to stay at comfortable temperature) and the external signals (it is hot outside) changes the importance of the visual signals. This implies that the weight of evidence should be modulated by its behavioral significance or salience, and not only by the uncertainty of its information source." (p. 519)

In creatures like human beings, deep contextualisation of bodily states can make distal or abstract outcomes salient—like conditions of autonomic arousal being inferred to be anxiety about an exam next week. A growing number of researchers seek to ground selfhood and emotion in interoceptive processes, particularly in their functional relation to allostatic regulation (Seth, 2015; Seth & Friston, 2016; Barrett & Simmons, 2015; Pezzulo, 2015). Crucially, because interoceptive inference largely involves predictive modelling of key physiological variables, it is apt to put greater emphasis on *control* over *discovery* (Seth & Friston, 2016). Self-evidencing most fundamentally involves allostatic control (Hohwy, 2016; Pezzulo, 2015). This results in the "a priori hyper-precision of visceral channels" (Allen & Tsakiris, 2018, p.7), in which interoceptive signals are assigned very high precision in virtue of communicating information about key homeostatic variables (Seth, 2014).

In the aim of successfully navigating the world over longer timescales, and selecting policies that result in survival—and not dispersion or non-existence—organisms must possess models of the future; in other words, they require *deep temporal models* (Friston et al., 2018). The generative models that endow organisms with the capability of inferring the consequences of future actions must have the property of *temporal thickness* (Friston, 2018). Temporal thickness allows the organism to anticipate the consequences of future actions, which confers the ability to select policies or action scripts that are favourable to the organism's continued existence. The minimisation of surprise through active inference on the FEP involves acting so as to reduce uncertainty, and to do this the system must model itself across time and counterfactuals:

> "...because active inference is necessarily system-centric the self-evidencing of motile creatures can only be elevated to self-consciousness if, and only if, they model the consequences of their actions" (Friston, 2018, p. 5).

Self-modelling emerges as a natural consequence of prospective action selection (Friston, 2018).
The self-model spans these multiple levels to track allostatic imperatives on different timescales. While the more basic aspects of the self (and generally the most phylogenetically ancient) aspects of the self are the fastest changing – such as being hungry – the higher levels of the self-model, in tracking the self on a much longer timescale, are "increasingly abstract, complex, and invariant; i.e., these high-level self-representations will be less likely to be affected by prediction error." (Friston & Limonowski, 2018). It's worth emphasising, however, that even the highest levels are fundamentally

interoceptively grounded. For instance, the motivation to sign up to a pension scheme to someone just entering the workforce can be understood to be underpinned by interoceptive inference on afferent interoceptive signals – manifesting, for example, in a feeling of anxiety when thinking about a future without a pension. The function of the self-model on this picture, then, is to flag motivational relevance on different levels of abstraction, were low-levels track immediate goals and pressing allostatic needs and high-motivate long-term goals and underpin a sense of personal identity.

Conceiving of the self through an active inference framework, the self-model functions to guide policy selection over various timescales in service of minimising expected free energy. Having a deep temporal model, and a corresponding counterfactually rich self-model, attunes organisms to the world and its affordances at different timescales. On this picture, narrative dimensions of the self at higher-levels constrain the self at lower levels in that the self-model "actively shapes itself over time to align with those higher level regularities" (Hohwy & Michael, 2017) where the narrative theory of the self constrain the planning and decision-making of the agent (Menary, 2008). 'Deep' self-models, that is, those that are hierarchically contextualised and 'temporally thick', allow for precision to be assigned according to goals on many different timescales. Inferred states of the self, on multiple timescales, inform the allocation of precision and what is 'salient' in the environment: "salience is literally defined by whatever has the most (or least) impact on visceral and autonomic homeostasis" (Allen & Tsakiris, 2018)—where highest hierarchical levels anticipate downstream consequences of actions and select policies accordingly (Friston et al., 2010; Pezzulo, 2015).

The functional role of the self-model, then, fits in with approaches such as the *affordance competition hypothesis* (Cisek, 2007; Cisek & Kalaska, 2010; Pezzulo et al., 2018), where different affordances jostle for precedence and are selected on the expected desirability of their outcome. Affordances that are salient for action depend on the inferred state of the self, manifesting as feelings – e.g. hungry, sociable, etc. The allocation of precision, "which confers salience on attended representations" (Friston et al., 2012), via dopaminergic gating is thus shown as crucially assigned relative to inferences about states of the self. This means the world appears different, and opportunities for action appear different, depending on inferences about the self—consider, for instance, how different an open dance floor appears to someone feeling confident compared to someone feeling shy.

A unified self-model allows for high-levels of the self-model to entrain salience at lower levels – for instances, an email regarding funding a funding opportunity might be very salient given longer term goals. A unified self-model also allows for different levels of the self-model to conflict, and actions to be prioritised accordingly—for instance, delaying the motivation to satisfy hunger when in a restaurant by going through the usual motions of calling the waiter rather than eating off another diner's plate. By integrating motivational salience across levels to find policies with the least expected free energy, the self-model assigns precision and salience to the world by weighing motivations on different timescales against each other.

This understanding of phenomenal self-modelling, and its underlying computational mechanisms, can be used to account for a wide range of disruptions in self-consciousness, such as drug-induced ego-dissolution (Letheby & Gerrans, 2017; Millière, 2017); meditation (Laukkonen & Slagter, 2020); and depersonalization (Gerrans, 2019). This thesis draws on disruptions in selfhood and consciousness in order to unpick both the relationship between consciousness and the self on this account, and to demonstrate how normal consciousness is structured and shaped by self-modelling mechanisms—even in cases of radical disruptions on self-consciousness. Disruptions in self-consciousness can also put pressure on particular philosophical claims, such as the claim that self-consciousness is necessary for phenomenal consciousness, rather than simply pervasive in normal experience.

In chapters 3 and 4, I give an active inference account of a few disruptions in self-consciousness—including psychedelic-induced ego-dissolution, meditation, and depersonalisation disorder. I subsequently use these accounts in chapter 5, to bring into focus why self-modelling should be understood as self-modelling should be understood as constitutive of the conscious condition. This argument follows a number of authors who posit that there is such a thing as a basic sense of self or self-consciousness in the background of any phenomenally conscious experience (e.g. Chalmers, 1996; Damasio, 1999), echoing William James' intuition that "whatever I may be thinking of, I am always at the same time more or less aware of myself" (James, 1961, p. 42).

This thesis can be divided into two primary parts. The first part (Chapters 1-4) introduces the PP framework and seeks to tie fundamental and pervasive features of conscious experience—such as the sense of self, the sense of agency, and affectivity and emotion—into the predictive processing framework. Locating consciousness within the predictive processing framework constitutes the second part of this thesis. I argue in favour of a subjectivity theory of consciousness within the active inference framework. That is, the self-modelling mechanisms posited in chapters 3 and 4 of the thesis I take to be constitutive of the conscious condition. A natural objection that flows from this is that both chapters 3 and 4 account for how self-consciousness can break down, while consciousness itself remains in tact. This seems suggestive that while the predictive processing approach to self-modelling may account for self-consciousness as *content* of consciousness, states where this self-conscious content is absent and consciousness is in tact indicates that self-modelling should not be considered as constitutive of the conscious condition. In chapter 5, I consider this question in detail and argue that consciousness itself can be understood in terms of fundamentally affective computational self-modelling mechanisms.

## Organisation

In **Chapter 1**, *Wilding the Predictive Brain,* I give an overview of the framework. The key idea of this chapter is to unpack the view of the brain as a hierarchical prediction machine. In particular, the focus is building up from the more simple formulations of perceptual inference to richer conceptions of prediction and the active inference framework. This is done by first considering simple perceptual inference, as inference about the hidden causes of sensation. The inferential nature of perception is a view dating back to Helmholtz. The active inference framework takes action to be part of the very same mechanism, where actions realise predictions, and perception and action turn out to be two sides of the same prediction error minimisation mechanism.

The next step, which proves crucial for understanding the phenomenology of selfhood in the chapters to come, is embodied active inference. Here, cybernetic and control theoretic principles are incorporated into basic active inference story, such that the organism acts to realise a certain state of the body or interernal milieu— grounding perception and action in the self-organisation principles thought to inhere in living things on the free energy principle.

**Chapter 2**, *Getting Warmer,* gives an overview of affective inference within predictive processing. The key concepts here are that affective inference is a kind of interoceptive inference. Emotions are thus understood in terms of predictive models about afferent inputs from the body. Importantly, the models of this input, just like with exteroceptive perception, are highly sensitive to context—connecting the predictive processing approaches to emotion to earlier appraisal theories.

**Chapter 3**, *Dissolving the Self,* introduces a core idea in the thesis, that phenomenal selfhood is underpinned by an *allostatic control model.* The idea here is that the sense of being a self or an agent is underpinned by an inference about endogenous control of the temporally deep consequences of action, and specifically, sensation related to self-evidencing outcomes. Affective inference, then, is understood as a particular kind of self-related inference, tracking broadly "how well am I self-evidencing?"

Having built an account of 'pre-reflective' selfhood in terms of allostatic control, I then offer an account of how the normal sense of self is disrupted in the case of psychedelic-induced ego-dissolution. To do this, I first build on Carhart-Harris & Friston's (2020) 'REBUS' model of psychedelic action as relaxing high-level priors. I argue that the phenomenological effects of psychedelics—'tripping'—can be understood in terms cast this in terms of a high Bayesian learning rate on sensory evidence, whereby prediction errors from across the cortex are afforded high precision. The result of this is that the system cycles through candidate hypotheses to explain away in the incoming sensory prediction error.

Based on this characaterisation of the action of psychedelics, I identify computational mechanisms underpinning psychedelic-induced ego-dissolution . The core idea here is that the normal sense of self critically functions to attenuate sensory evidence. The failure of this sensory attenuation results in a breakdown in the mechanisms of attribution of self and non-self, as the system ceases to posit itself as the endogenous controller of sensation.

I then give an account of the affective tone of the experience. Ego dissolution is characteristically ecstatic, but can sometimes be very scary and distressing. I account for the ecstatic nature of ego-dissolution as arising from a relaxation of high-level prior preferences in the generative model. Dysphoric ego-dissolution experienced in 'bad trips', by contrast is accounted for by high endogenous precision on the preference to retain control over the experience, where the control is increasingly failing as action outcome contingencies become increasingly unpredictable.

**Chapter 4,** *Losing Ourselves,* compares the selfless states of sufferers of depersonalisation disorder with selfless states experienced by meditation practitioners. This chapter builds on the allostatic control model outlined in the previous chapter. There is particular focus here in conceptual separation of agentive control—the agent's model of the expected sensory consequences of action, and how the sense of self is grounded in inference about this control over prior preferences.

Depersonalisation disorder is a dissociative disorder that is characterised by feeling like one doesn't really exist, or is somehow disconnected from themselves and reality. In this state, the sufferer's memories and perceptions can seem as if they belong to another person, and they may not feel like they are the author or controller of their lives, sometimes describing feeling like an 'automaton'. States of deep meditation are also known to generate 'selfless' experiences of ego-dissolution. By contrast, these experiences are associated with pervasive joy and euphoria. Indeed, these states of selflessness, in many traditions such as Buddhism, are highly sought after and identified with 'enlightenment'. This chapter applies the computational description of the sense of self account to account for the phenomenology associated with depersonalization and deep meditation practice.

In **Chapter 5**, *Consciousness in active inference*, I argue that phenomenal self-modelling can act as the foundations for a theory of consciousness. At this point in the thesis, I have argued that the predictive processing framework can account not only for perceptual contents, but is also illuminating the computational underpinnings of the sense of self, and how the disruptions of self-consciousness can be accounted for in terms of an inference about allostatic control.

The question remains whether predictive processing can provide an account of consciousness itself. One approach to consciousness within predictive processing is to identify consciousness with the self-modelling mechanisms that inhere within the active inference framework. Friston (2018) takes this approach. However, numerous commentators have highlighted a problem with so-called 'subjectivity theories'—theories of consciousness that identify consciousness with self-consciousness, namely that disruptions in self-consciousness, while consciousness remains intact, is suggestive that self-consciousness is not necessary nor constitutive of consciousness itself. I then use these cases to show that consciousness is understood as inherently affective, understood to be ‘shaped’ or structure by the hierarchy of self-models that constitute the ‘lens’ of perception—that is, filter sensory inputs based on the *meaning* they have for the organism.

Chapter 6, *Expecting Action*, bridges this account of consciousness in active inference with Ward, Roberts & Clark’s (2011) conception of the contents of visual perceptual experience, understood in terms of a ‘poise’ over an action space’. This brings into focus how the account put forward in this thesis can be understood to be both embodied and action-oriented. The action space account understands visual experience as infused with the possibilities the agent has for interacting with the environment; “for pursuing and accomplishing one’s intentional actions, goals and projects” (Ward et al, p 383). In this chapter, the compatibility of this account with the allostatic control model is brought to the fore, where the active inference approach to consciousness argued for in the previous chapter can be cast in terms of poise over an action space. The account is then generalized to a *cognitive* action space, where the agent has “poise” over opportunities for mental actions. Building on this view, the chapter closes with some reflections on how creatures which the capacity for mental action may come to puzzle on their own consciousness as a mysterious phenomenon.

## Author contributions

The chapters in this dissertation have either been published in article form or have been submitted to journals. As a result, there is overlap across different chapters—for instance mechanics of active inference are described several times. Below I provide, if existing, the reference to the published version, and if co-authored I note the relative author contributions.

**Chapter 1** has previously been published as:
Nave, K., Deane, G., Miller, M., & Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(6), e1542.
**Author contribution:** Equal contribution—all four authors conceived and co-wrote the paper together.

**Chapter 2** has previously been published as:
Wilkinson, S., Deane, G., Nave, K., & Clark, A. (2019). Getting warmer: predictive processing and the nature of emotion. In *The value of emotions for knowledge* (pp. 101-119). Palgrave Macmillan, Cham.
**Author contribution:** Equal contribution—all four authors conceived and co-wrote the paper together.

**Chapter 3** has been previously published as:
Deane, G. (2020). Dissolving the self. *Philosophy and the Mind Sciences*, *1*(I), 1-27.
**Author contribution:** Conceived and written by myself.

**Chapter 4** has been previously published as:
Deane, G., Miller, M. D., & Wilkinson, S. (2020). Losing Ourselves: Active Inference, Depersonalization and Meditation. *Frontiers in Psychology*, *11*, 2893.
**Author contribution:** Primarily conceived and co-written by myself and Mark Miller. Contributions and revisions by Sam Wilkinson.

**Chapter 5** has been revised and published as:

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.

**Author contributions:** Conceived and written by myself.

**Chapter 6:** *Expecting Action*: Predictive Processing and the Construction of Conscious Experience

Submitted to a journal. Conceived and written by all four authors.

# Chapter 1: Wilding the Predictive Brain

The Predictive Processing (PP) framework casts the brain as a probabilistic prediction engine that continually generates predictions of the causal structure of the world in order to construct for itself, from the top down, incoming sensory signals. Conceiving of the brain in this way has yielded incredible explanatory power, offering what many believe to be our first glimpse at a unified theory of the mind. In this chapter, the picture of the mind brought into view by predictive processing theories is shown to be embodied, deeply affective and nicely poised for cognitive extension. We begin by giving an overview of the main themes of the framework, and situating this approach within embodied cognitive science. We show perception, action, homeostatic regulation and emotion to be underpinned by the very same predictive machinery. We conclude by showing how predictive minds will increasingly be understood as deeply interwoven with, and perhaps extended into, the surrounding social, cultural and technological landscape.

## 1. Introduction

Recently a new perspective dominates many discussions of mind and cognition. That perspective depicts the brain as essentially a probabilistic prediction engine, dedicated to the task of minimizing the disparity between how it expects (predicts) the world to be and the evidence presented by the sensory flow. Part of the power of the framework lies in the elegant suggestion that much of what we take to be central to human intelligence - perception, action, attention, emotion, learning and language - can be understood within a simple framework of predictions and error reduction. In what follows we will refer to this general approach to understanding the mind and brain as *predictive processing* (PP).

While the predictive processing framework is in many ways revolutionary, it can appear to commit proponents to a traditional 'neurocentric' stance concerning the mind. Such a stance depicts the mind as, in essence, what brains do, and it has tended to downplay the contributions of body, world, and action.  However, a closer look reveals a much richer picture. Today, many researchers are exploring views of the predictive brain that allow for the body and the surrounding environment to make robust contributions to the predictive process itself. These recent developments strongly suggest that while it's true that predictive models can get us a long way in making sense of what drives the neural-economy, a complete picture of human intelligence requires us also to explore the many ways that a predictive brain is embodied in a living body, permeated with affect and embedded in an empowering socio-cultural niche. In this chapter, we explore this promising evolution (see table 1).

| **Evolution of the framework** | **Key readings** |
|---|---|
| Perceptual Inference | Helmholtz (1867/1910); Gregory (1997); Lee and Mumford (2003); Rao and Ballard 1999 |
| Simple Active Inference | Friston, Adams et al 2012; Clark 2013, 2016 ; Friston et al. 2015, 2016, 2017. |
| Embodied Active Inference | Allen & Friston 2018; Gallagher & Allen 2018; Kirchhoff 2018; Bruineberg et al. 2018; Rietveld & Bruineberg 2017; Allen & Tsakiris 2019; Allen & Friston, 2018; Linson et al. 2019; Gallagher 2018; Seth 2013; Barrett & Simmons 2015; Joffily & Coricelli 2013; Van de Cruys 2017; Kiverstien et al. 2017; Hesp et al. 2019. |
| Extended Predictive Minds | Kirchhoff & Kiverstein 2019; Constant et al. 2019; |

**Table 1**: Evolution of the Predictive Processing Framework

We start by introducing the predictive processing framework itself. We go on to outline how different camps within predictive processing research are viewing its amenability to embodied cognitive science, and lay some conceptual groundwork for ways to think about the predictive mind as active and embodied. We then turn to recent work on feelings and emotions, and review a new take on affect value and its role in directing core estimations of the reliability and value of prediction errors. We conclude by showing how such an active, embodied and affect-laden predictive organization would be well-poised to extend cognitive processing into the social, cultural and technological landscape.

## 2. Perceptual Inference

An increasingly popular theory in philosophy and neuroscience is that the perceiving brain is fundamentally an engine of prediction (Bubic et al. 2010; Friston 2010; Howhy 2013; Clark 2013). This engine, so the story goes, makes sense of the world by actively generating predictions of the incoming sensory stream. According to this view, much of what we take to be central to human intelligence - perception, action, attention, cognition, emotion, learning and language - are all underpinned by this common predictive mechanism (Clark 2013: 181). In this chapter, we assume that some version of the predictive processing story has merit - for a balanced review of the state of the evidence, see Walsh et al. (2020).

This bedrock 'predictive processing' (PP) story has roots in influential accounts of perception (Kant 1781/1929; Helmholtz 1867/1910; Gregory 1997). How does the brain unearth, and enable us to perceive, a complex world of objects and events given the many challenges posed by noise and ambiguity? This challenge has been referred to as the 'problem of perception' (Hohwy 2013). To bring this into focus, consider the fact that any object can induce a vast variety of sensory patterns: the same object can be encountered from different angles or in changing atmospheric conditions. Moreover, many different objects produce similar sensory signals: a picture of an object and the object itself, or a partially obscured object and a fragmented one (e.g. a cat walking behind a picket fence). To make matters worse, we must also be able to explain how the brain is able to separate out the salient information from the unimportant sensory noise. The PP model 'earns its salt' by offering an elegant and unified solution to such challenges.

In extremely brief terms, the predictive brain gets a grip on the noisy, ambiguous world presented by the sensory flow by continually learning about the vast network of temporal-spatial regularities that reflect, and in some ways constitute, its environment. Our world is filled with learnable regularities - both natural and synthetic, ranging from the very fast (e.g. patterns in a swiftly moving river) to very slow (e.g. the river eroding the land around it). Through tracking fast regularities we grasp the fine-grained details about a scene. Tracking slower regularities provide the kind of abstracted and enduring information (for example, about the unchanging identity of the river) that can help contextualize and constrain more local predictions, and informs predictions of how states of affairs are likely to evolve on longer timescales.

The predictive brain uses its knowledge of regularities and patterns to make increasingly refined predictions about what objects and events are most likely to be responsible for the signals it receives from the environment. Perceptual experience, then, is the top-down "best guess" at the hidden causes of incoming sensory signals. Where there is a significant mismatch between prediction and incoming signals, the discrepancy (the residual error, or "prediction error")  moves forward (or "up") through the hierarchical system helping to refine predictions or (see next section) to recruit actions aimed at making the sensory stream fit proprioceptive predictions, or at improving the state of information.

The prediction error minimizing routine is modulated at every level by a set of second-order expectations that track the reliability, or inverse variance, of the predictive system's own estimates given the state of the organism and the current context. It uses this estimation of reliability (referred to as 'precision weighting') to flexibly adjust the gain (like turning the volume up) on particular error units . Error units carry all the unexplained sensory information, and increasing their weighting increases the impact that information has on the unfolding process. This allows the system to flexibly modify the degree to which it relies on incoming prediction errors from the sensory periphery or prior beliefs about the state of the world (see Friston 2009, 2010). For example, while listening to your favourite song in the shower, it would be useful to turn down the influence on the sensory signals produced by the flowing water and rely more on our clear memories of the song (Clark 2016: 92).

The novel addition this theory makes to traditional, feedforward-dominated perception research is that perception is not explained by incoming signals alone, but crucially also includes active  top-down predictions about their shape, reliability and what they could mean. Perception thus constructed is sometimes said to involve a kind of "controlled hallucination" – what we see is impacted by what we predict ought to be out there. Of course what we predict is itself continually tuned by the actual sensory signals, which works (in normal functioning brains) to anchor those predictions to reality. Nonetheless, the hard perceptual work is here accomplished mostly by the internal model (the 'generative model') that is constructing the predictions, leaving the incoming sensory information the task of (in effect) critiquing those predictions until a better fit is achieved.

For example, Anthony Norcia's Coffer Illusion (Fig. 1) is usually first perceived as a grid of squares, like wooden panels on a door. However, once prompted to search for the circles, our subpersonal predictions about the content of the image, and so the corresponding experience of it, shifts dramatically. This despite the fact that the incoming pattern of sensory stimulation has not altered at all.

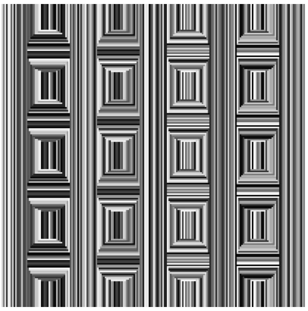

Fig.1 The Coffer Illusion, (Norcia, 2006)

This approach openly opposes classical feedforward-dominated perceptual models. In the not too distant past the brain was commonly characterized as a relatively passive organ. Dormant neurons were thought to patiently await incoming signals to jolt them into action. When signals did arrive, they were thought to roll in from the sensorium and flow upward through the neural hierarchy increasing in complexity along the way (Marr 1982). In direct contrast, PP describes the brain as fundamentally proactive - the brain actively generates perceptions by continually attempting to recreate from the top-down the world of sensory signals.

## 3. Simple Active Inference

If we think primarily about perception alone, it can seem as if the fundamental goal of the predictive brain is to reconstruct the distal environment on the basis of noisy and ambiguous sensory information. But as noted above, PP places action at the heart of the prediction error minimizing process too. This opens the door to a non-reconstructivist understanding that is more closely aligned with work on the embodied mind (Varela et al. 2016; Chemero 2011). Such approaches reject the characterisation of our perceptual goals as reconstructive in the first place. As a classic statement of the tradition, cited by Clark (2016) puts it:

> "The overall concern of an enactive approach to perception is not to determine how some perceiver-independent world is to be recovered; it is, rather, to determine the common principles or lawful linkages between sensory and motor systems that explain how action can be perceptually-guided in a perceiver-dependent world." (Varela et al 1991, p.173)

Perhaps the active PP system can make do with something other than a faithful reconstruction of the distal environment? To take a well-worn but still useful example, Phillip Fink and colleagues (2009) show that a baseball outfielder need not first model the entire onward trajectory of the baseball relative to their position, and to the field, in order to then begin the act of moving to catch it. All that is needed is an ongoing coordination strategy called 'Optical Acceleration Cancellation' (OAC) – that involves moving such that the ball stays at a stable position in the retinal field, until it is close enough to catch. In order to successfully execute this strategy an agent requires no internal physics engine, no knowledge of aerodynamic equations governing the flight of a sound, slightly irregular, projectile in a mild North-Westerly wind. All they require is an understanding of the lawlike relations between their motor output and the position of the ball's projection on their retina.

In PP terms, as Clark (2016) explains, this becomes a matter of assigning high-precision weighting to errors related to the prediction that the optical projection of the ball remains at a stable location on the retina. In such a way the rest of the system's actions are recruited around the quashing of this particular error signal, to the neglect of most else happening on the field, until the desired state of catching the ball (or the undesired state of colliding with a teammate employing the same strategy) is reached. Here there is no prior process of tinkering at the generative model until a lack of overall error provides adequate comfort that we've formed an accurate representation of the external world, and action may now begin. Rather, successful action is itself the ongoing control of a small portion of the sensory flux within those constraints that the system predicts will lead towards its target state.

As Anil Seth (2015) suggests, we can think of this "non-reconstructivist" approach to PP as offering a mechanistic rendition of earlier sensorimotor theories of perception (O'Regan & Noe 2001). Such 'fast and frugal' strategies are much more suited towards the ongoing guidance of an organism that must constantly keep afloat in a fast-changing environment (Clark 2016). For Clark, the availability of locally-effective non-representational strategies is not an argument that we should abandon all representation talk however. Rather the strength of active PP is the offer of, "a systematic way of combining deep, model-based flexibility with the use of multiple, fast, efficient, environmentally exploitative, routes to action and response"(2015, p18). In order for the PP system to effectively deploy such 'fast and frugal' strategies as OAC, it must also be able to monitor slower-changing contextual factors (such as whether one is actually engaged in a game of baseball, or merely a participant) in order to ascertain when the circumstances are ripe for their deployment. This is why the PP system requires hierarchical depth, such that high-level states may target these large-scale increasingly invariant patterns throughout the fast fluctuations of the sensory stream.

Unlike on the reconstructivist story, these high-level action-oriented representations do not allow us to "throw away the world" (Clark 1999) when we engage in planning our next action, but rather coordinate our interactions with the world at multiple levels of spatiotemporal grain. Nor is the correctness of these action-oriented representations contingent upon the rejection of any sceptical hypotheses. If a current affordance for ball-catching-action is correctly detected, then deploying OAC will guide the evolution of the skilled outfielder's sensorimotor interactions to the target ball-in-hand state. This model of current sensorimotor contingencies fulfills its purpose successfully – and it does so, we should also note, irrespective of whether the hidden causes interacting with our sensorimotor array were instantiated by mischievous demons, curious scientists, or strange and charming fundamental particles.

## 4. Embodied Active Inference

If not the goal of representational fidelity, then what is the end towards which our sensorimotor interactions are being coordinated? Within the recent literature on PP, we can distinguish another (even more radically embodied) strand that treats the active, embodied brain in ecological and enactive terms (Allen & Friston 2018; Gallagher & Allen 2018; Kirchhoff 2018; Bruineberg et al. 2018).

Enactivists subscribe to a mind-life continuity thesis that takes cognitive processes and living processes to work according to the same fundamental organizing principles with the shared imperative of maintaining the organism's integrity through adaptive regulation of environmental interactions (Varela et al. 1991; Thompson 2007; Di Paolo 2005). In working to produce, sustain and conserve its identity over time in its interactions with the environment, the organism enacts or “brings forth” a meaningful world, a process termed “sense-making”. Cognitive and affective processes work together in this framework to steer the organism through the world in pursuit of what is significant. Colombetti writes that “Cognition from an enactive perspective is, rather, the capacity to enact or bring forth a world of sense, namely, an Umwelt that has a special significance for the organism enacting it… cognition as sense making entails that cognition is simultaneously also affective.” (Colombetti 2014, p.18).

In this ecological-enactive rendition of PP, affect and cognition unfold in ways that are deeply reciprocally interactive with the whole animal-environment system. Environmental surroundings, meanwhile, are conceived in relation to the affordances or action possibilities they offer to the organism. A familiar illustration that effectively captures this logic in its simplest form is that of bacterial chemotaxis. As Varela (1991) describes, sugar is necessary to fuel the metabolic processes responsible for the ongoing production of the bacterium’s body,  thus sugar gains affective significance in relation to the bacterium. And because sugar is not, typically, evenly distributed throughout a solution, so a bacterium must engage in adaptive actions in order to seek out and move towards increasing concentrations of it. If it does not, it will not remain a bacterium much longer. The nutritional value of sugar and the imperative of moving towards it has no pre-existing validity independent of the bacterium, but rather is brought about by the relation between its needs and activities, and the sugar-seeking affordances of its environment.

As we have noted, cognition is typically separated from affect in the philosophical and psychological literature on emotion. Appraisal and evaluation are associated with cognition, while affect is identified with changes in the autonomic system in the body. Appraisal according to this ecological-enactive approach is instead understood as involving the whole living body of the organism (as it prepares to act on relevant affordances). Enactivists theorists of PP sometimes refer to this using the French-phenomenologist Maurice Merleau-Ponty's notion of a ‘tendency towards an optimal grip’ (Rietveld & Bruineberg 2017). ‘Grip’ here refers to the organism's bodily stance in relation to its current situation. We use this term because grip, understood as bodily readiness for apt action, is something the organism must actively maintain in relationship with the changing environment. Prediction errors then signal an increase in disequilibrium in the organism-environment system as a whole - for instance, declining sugar levels, which will negatively impact readiness for action. Such disequilibria reflects a divergence in the sensory states (exteroceptive and interoceptive) the organism expects to occupy given the kind of organism it is, its current state and the niche it inhabits. The organism then acts to reduce this disequilibrium or to make it the case that it comes to occupy the sensory states that it expects to be in (given the life it leads). In other words, it acts so as to stay within the window of viability that defines it as the kind of being it is.

In this process, gross action is simply one component. Of equal importance is what might be thought of as 'inner action'. Indeed, it has long been argued that the reason for having a brain is to keep organisms in 'continuous equilibrium' with their environment (Pavlov 1927), and as part of this process of efficient management of metabolic resources to drive survival and reproduction has been described as the 'core task' of all brains (Sterling & Laughlin 2015). In order to stay alive and viable, an organism needs to keep its bodily states within certain bounds. For instance, a body temperature of 40 degrees centigrade is not conducive to continued existence for a human being. Regulating these 'essential variables' – homeostasis – requires sensing the global physiological conditions of the body through afferent signalling to the brain of the internal state of one's body. Sensing of the global physiological conditions of the body – such as cardiac signals, states of the gut and viscera, air supply and glucose plasma levels – is known as interoception (Craig 2003). Building a predictive model of the state of the body, via the very same inferential mechanisms underpinning perceptual inference, allows the brain to engage autonomic action to bring the body into homeostatic balance. For instance, a hyperthermic animal can bring its body temperature into viable bounds through engaging autonomic reflexes such as perspiration. In order to stay viable on longer timescales, however, it will need to act – for instance it could move to a cooler place, like a shaded area under a tree. These actions – those that allow the brain to regulate the state of the body – are called allostasis. To stay viable on longer timescales organisms need to take prospective actions, anticipating dyshomeostatic conditions before they arise and acting to avoid them.

Enactivist predictive processing regards this fundamental imperative towards continued existence via homeostasis and allostasis a kind of "first prior" (Allen & Tsakiris 2019). In other words, "[t]he brain is in the game of predicting the world, but only as a means to the end of embodied self-preservation" (Allen & Friston 2018, p.12). This means that the interoceptive signals, in tracking key homeostatic variables, are deemed highly precise - they are given "a priori hyper-precision of visceral channels" (Allen & Tsakiris 2018). Unlike in perceptual inference, where beliefs are hypotheses can be much more malleably shaped so as to fit the world, interoceptive inference tracks physiological variables which must be maintained within quite narrow bounds (Seth & Friston 2016). This is intuitive – the system that simply updates its perception of its body temperature to 50 degrees centigrade, rather than acting to cool down (either through autonomic reflexes such as by perspiring, or allostatically by moving to a colder place) is unlikely to stay alive for long.

These visceromotor channels, then, are highly precise and resist simple revision in favour of action – that is, they minimize prediction errors given evolutionarily endowed prior expectations that the organism will continue to exist. . But complex adaptive strategies also require prospective control, that seeks to minimize whole trajectories of error that reach far into the future. This demands deep *temporal models,* and recasts planning as a problem of probabilistic inference (Botvinick & Toussaint 2012). On this view, action selection is based on prior beliefs about the expected consequences of a course of action or sequence of actions, referred to in the literature as an action 'policy' (Friston et al. 2015, 2017). Expected prediction error is the error the agent predicts were it to pursue a particular action policy, given knowledge of the contingencies in the world, and knowledge of itself (such as the metabolic resources it has available to complete the action).

Intuitively, some actions are more likely than others to lead to desirable or 'expected' outcomes – such as leaving a building through a ground floor door as opposed to a 6th floor window. Selection of actions that minimize expected error over longer timescales, thereby maximizing the probability of long-term existence here rests on a balance between the pragmatic and epistemic affordances of action. The pragmatic affordances of action – such as having a drink on a hot day – are readily apparent in terms of the drive to maintain homeostasis. Epistemic actions, by contrast, refer to actions that improve our state of information (Kirsh & Maglio 1994;  Kiverstein et al. 2017). It is here that curiosity and exploratory behavior find a home within an active inference framework (Friston et al. 2017; Kaplan & Friston 2018; Kiverstien et al. 2017) as they enable organisms to increase their predictive grip in the long-term, by improving information and reducing key uncertainties (Friston et al. 2015, 2016, 2017). We return to these issues when we look at 'extended predictive minds' in section 5 below.

## 5. Emotions, Feelings, and Error Dynamics

The homeostatic perspective has also been invoked in thinking about how the PP framework may be extended to discussions about feelings, emotions and moods (Ainley et al. 2016; Allen et al. 2016;

Apps & Tsakiris 2014; Barrett & Simmons 2015; Kiverstien et al. 2020; Bruineberg & Rietveld 2014; Clark 2016; Gu et al. 2013; Seth 2013, 2014; Seth et al. 2012). Anil Seth, for example, proposes that PP applies just as neatly to interoception as it does exteroception (2013). Interoception refers to our internal sense of physiological changes in the body – somatic, visceral, vascular, and motor (Craig 2002, 2003). According to Seth, top-down predictions about the source of interoceptive signals counter-flow with bottom-up interoceptive prediction errors (just as in exteroceptive experiences). Feelings arise from the ongoing integration of these various predictive representations.

Inward-looking error signals are hypothesized to be minimized in an analogous way to sensorimotor predictions: either the error modifies the model to fit the inner world, or autonomic reflexes are initiated which influence the body to fit the prediction. In this model autonomic reflexes are called on to fulfill interoceptive predictions. Consider the example of hunger: when blood sugar levels drop below expected levels interoceptive error is generated. These errors update top-down expectations leading to the subjective experiences of hunger. While the interoceptive error is explained away at one level of the hierarchy as 'hunger', the hunger state itself produces error that is s resolved via autonomic feedback loops that metabolize fats and/or feedback loops that establish and guide allostatic action sequences that lead to finding and consuming sugary foods.

This predictive approach to explaining bodily feelings suggests new possible dimensions to classical somatic theories of emotion. For example, William James defended a view of emotion as the perception of the physiological changes that result from an exciting encounter (1884). 'Fear', for James, was thought to be constructed by our interoceptive perceptions of the internal bodily changes that are characteristic of fear (e.g. sweating, intercostal tightening, etc.). Subjectively speaking,  interoceptive awareness manifests as a diverse set of feelings including those of "pain, temperature, itch, sensual touch, muscular and visceral sensations...hunger, thirst, and "air hunger" (Craig 2003 p. 500). The feeling of fear, if James is right, is thus essentially the detection of an interoceptive physiological signature that has already been induced by exposure to the threatening situation.  In other words, we do not shake because we are afraid of the angry dog, we *shake* and that shaking makes up our fear.

The trouble with such a simple story is that it suggests a one-to-one mapping between distinct emotional states and distinctive 'brute-physiological' signatures, and no such mappings have been found (Critchley et al. 2004). To remedy this, Seth (2013) and Pezzulo (2014) suggest we may integrate basic information (e.g. about bodily arousal) with higher-level predictions of probable causes. This would help explain why the sensation of a fast-beating heart feels so different when we suspect heart-failure than it does after (say) a bout of vigorous exercise. It also helps explain the kinds of vicious cycle seen in panic attacks, when the best guess (heart attack) itself impacts bodily response, delivering spurious 'evidence' in favour of the heart-attack guess itself.

PP offers an elegant and believable neurocomputational explanation of how these recursive processes might unfold, fueled by the drive to continually reduce the error between top-down predictions and bottom-up interoceptive signals (see also Clark 2016; Barrett & Simmons 2015; Seth & Friston 2016). As Seth notes, these two directions of interaction - from prediction to bodily changes, and from bodily changes to prediction - unfolds "continuously and simultaneously underlining a deep continuity between perception and action" (Seth 2013: 566). The brain's best guess at the cause of some perceptual signal activates somatic patterns (see also Damasio 1994) that prepare the body to respond appropriately *and* help the system predict what will happen next. The reactivation includes both autonomic changes and explicit actions (including gestures, facial expressions, postural changes, etc.). These bodily changes provide the basis for the next wave of interoceptive information to be integrated and matched against the evolving prediction (which includes exteroceptive information, memories and predictions), and it is these ongoing reciprocal exchanges that structure our emotional experience. Prediction and incoming signals co-evolve in cycles attempting to minimize discrepancies between model and signal, becoming accessible to conscious awareness as a best-guess stabilizes (Harrison et al. 2010; Craig 2002, 2009).

This provides a very natural way of accommodating large and long-standing bodies of experimental results showing that the character of our experience depends both on the interoception of brute bodily signals and higher-level 'cognitive appraisals' (for a review, see Critchley & Harrison 2013). But the emerging predictive processing account of qualitative experience should not be thought of as a 'two-factor' theory as such, but rather, the claim is that a single, highly flexible, inferential process fluidly and constantly combines top-down predictions with bottom-up sensory information. Subjective feeling states are then determined by the ongoing unfolding of this single process.

More recently a view of interoception has begun to take shape that depicts bodily feeling as rather more than just another stream of information for the predictive brain.. The idea is that affective changes in the organism's body might play a special, crucial role in the predictive process itself, through a tight relationship between affectivity and precision weighing.

To bring out this important relationship, we might consider recent work on *error dynamics* (see Kiverstein et al. 2017; Hesp et al. 2019; Kiverstien et al. 2020; Miller et al. 2020). 'Error dynamics' refers to the temporal comparison of error reduction rates (see Joffily & Coricelli 2013; Van de Cruys 2017). For a predictive organism to thrive it needs to be sensitive to error and error reduction, but also the *rate* at which errors are being reduced. If we think about error minimization happening at a certain speed over time, then the predictive organism must also be sensitive to changes in velocity - accelerations and decelerations in error minimization over time. This sensitivity indicates how well or poorly the organism is doing at reducing uncertainty. Error dynamics, as such, are second-order processes closely related to precision weighting[1]. In essence, precision is about the error bars on the current best-guesses, while information about the rate of error minimization (relative to expectations) is part of the mechanism that helps set these error bars, and that identifies actions and environments that look set to reward further exploitation. The positive and negative affective tone that accompanies experience may be the conscious reflections of these error dynamics.

Error dynamics, in other words, are made available to the system as emotional valence – feelings of pleasure and displeasure, or states of attraction or repulsion (Joffily & Coricelli 2013; Van de Cruys 2017).[2] When things are going well for the predictive agent (and its behaviours are resulting in a more certain future) it feels good. When it's struggling to get a handle on the scene, or is unable to manage the complexity of some task, it feels bad. To be clear, this feeling should not be seen as something over and above the tracking of error dynamics, but rather the feelings are a reflection of the quality of the organism's engagement with (or 'grip upon') the environment (see also Polani 2009). As Van de Cruys (2017) writes, "“Emotions … appear as the continuous non-conceptual feedback on evolving —increasing or decreasing— uncertainties relative to our predictions. The upshot of this view is that the various emotions, from "basic" ones to the non-typical ones such as

[1] For more discussion on this point see Kiverstein et al. 2017 and Miller et al. 2020.

[2] This notion of valence as emerging as a form of 'prediction error dynamics' has already found a home in both artificial intelligence and robotics circles (Schmidhuber 2010; Kaplan & Oudeyer 2007), and research on intrinsic rewards and adaptive behaviours in humans and non-humans (Kaplan & Oudeyer 2007).

humour, curiosity and aesthetic affects, can be shown to follow a single underlying logic".

There is then a conceptual connection between our sensitivity to our own error dynamics, feelings, and the 'tendency towards an optimal grip'. Feelings reflect the need to keep expected uncertainty to a minimum in our interactions with the environment. This means that the organism succeeds in improving its overall condition in relation to its environment, keeping in touch with what matters, in ways that reflect both bodily state and trajectories of error resolution. This folds embodied, environmentally embedded) value back into the heart of the predictive system[3].

## 6. Extended Predictive Minds?

Let's end by seeing how this all scales up to whole brain-body-world ecologies. For here too, the predictive brain plays a crucial and distinctive role. Brains like this will congenitally trade real world action against on-board computation (by using an app, for example) whenever that is estimated to be the best way to reduce key uncertainties (hence to minimize future errors) given current goals. Such 'epistemic' actions are chosen, we saw,  so as to minimize expected future prediction error. The point to notice is that epistemic actions selected to reduce expected uncertainty in this way can exploit any amount of reliable environmental structure and scaffolding. Inner (brain-based) strategies will thus emerge that rely heavily on the use of external structure and resources.

PP here solves, at least in principle, the so-called  'recruitment puzzle' concerning the 'extended mind' (Clark 2008). The puzzle concerns just how the canny cognizer manages to recruit, on the spot, whatever mix of problem-solving resources will yield an acceptable result with a minimum of effort. In Clark's 2008 treatment, the puzzle was described like this:

> "[our story] bequeaths a brand new set of puzzles. It invokes an ill-understood process of "recruitment" that soft-assembles a problem solving whole from a candidate pool that may include neural storage and processing routines, perceptual and motoric routines, external storage and operations, and a variety of self-stimulating cycles involving self-produced material scaffolding [e.g. sketching]. And at its most radical, it depicts that process as proceeding without the benefit of a central controller." (p. 137)

[3] Addiction then emerges as a case in which this tendency towards an optimal grip breaks down (Miller et al. 2020).

By minimizing estimated future prediction error, a PP system speaks to all these demands., The 'process  of recruitment' is simply the forward-looking use of estimated uncertainty (future prediction error) to select transient coalitions of internal and bio-external resources. In this selection process the enabling of internal resources (different brain areas, and different information from within the generative model) is accomplished by the very same means as the selection or enabling of external resources. The changing web of internal neuronal influence is selected using the very same means (variable precision weighting) and for the very same reasons (the reduction of estimated uncertainty) as the use of bio-external resources such as apps and notepads. The guiding principle binding all the cases is simply the energy-efficient minimization of expected future prediction error. And in place of a central controller, we now find only that rolling process of error minimization itself.

The wider world ends up looking somewhat different too. For we can now see much of the human-built world as itself a prime reservoir both of achieved precision estimations and of tools for cheaply estimating precisions on-the-fly. Think, for example, of the way we paint red lines beside roads, or fly red flags on dangerous beaches. These otherwise arbitrary structures attract attention and act as local proxies for precision – for our precision estimating brains (see Roepstorff et al. 2013). And as we behave in our present niche, we gradually alter it – the roads leading to that red flag flying beach may fall into disuse over time, and new ones emerge leading to safer swimming spots. In the ensuing dance between predictive brains and the forces of cultural and socio-technological change may lie the explanation for much that is distinctive about the human mind.

## 7. Conclusion

Predictive brains, we have argued, are beautifully positioned to weave thoughts, emotions, mental actions, bodily actions, and environmental opportunities into seamless webs that both serve and express our purposes. In this piece, we have displayed (Table 1) some of the recent evolution of these ideas. Simple appeals to generative models, error minimizing strategies and precision weighting get us a long way in making sense of what drives the neuroeconomy. But a complete picture of human cognition will require us to 'wild the predictive brain' by factoring in the many ways that the predictive brain operates as part of a wider system. This means locating the neural prediction machinery within an environmentally situated, active body, and appreciating the crucial place of affect in the predictive mind.

# Chapter 2: Getting Warmer: Predictive Processing and the Nature of Emotion

Predictive processing accounts of neural function view the brain as a kind of prediction machine that forms models of its environment in order to anticipate the upcoming stream of sensory stimulation. These models are then continuously updated in light of incoming error signals. Predictive processing has offered a powerful new perspective on cognition, action, and perception. In this chapter we apply the insights from predictive processing to the study of emotions. The upshot is a picture of emotion as inseparable from perception and cognition, and a key feature of the embodied mind.

## 1. Predictive Processing and Emotion – The Story So Far

Emotion and cognition are typically thought of in contrast to one another, sitting on opposite sides of a divide between passion and reason, the hot and the cold. But what does our best theory of the brain and central nervous system (CNS) tell us about the nature of emotion?

According to an increasingly popular framework in computational neuroscience, the brain is a hierarchically arranged prediction machine (Clark (2013a)). Contrary to once-popular feedforward approaches, the brain does not simply take inputs from the outside world, process them, and pass them deeper and deeper into the processing economy. Instead, whenever information from the world impacts on your sensory surfaces, it is already, even at the earliest stages, greeted by a downward-flowing prediction on the part of your nervous system. This prediction comes from your brain's best model of what is going on in the world, and this model is constantly being updated by the mistakes it makes, by the so-called 'prediction error signal', which it constantly tries to keep to a minimum (Lee and Mumford (2003), Rao and Ballard (1999)). In recent versions, this signal is weighted according to how reliable or salient the brain estimates the sensory information to be, relative to its best predictions. This 'precision-weighting' device operates at every level of processing. It implements attention, and allows us flexibly to balance top-down prediction and bottom-up sensory information (see Feldman and Friston (2010) and Clark (2013b))

The core business of brains like ours, if these stories are on track, is the minimization of precision-weighted errors in the prediction of sensory inputs (see Friston (2005) – and for comprehensive reviews, see Howhy (2013), Clark (2013a)). Importantly, the minimization of precision-weighted prediction error isn't always achieved by the brain updating its models of the world (which results in perception and belief). Instead it is sometimes achieved by bringing the world, usually the body, in line with the model (Feldman and Friston (2010), Clark (2016) Chapter 4). The result of this is bodily action.

According to early work in predictive processing (e.g Lee and Mumford (2003), Friston (2005)), what you *perceptually* experience is determined by the model that your brain adopts so as to best predict *exteroceptive* sensory signals such as incoming visual and auditory information. Building on this basic idea, it has recently been suggested (Seth (2013)) that what we *emotionally* experience is determined by the model that your brain adopts so as to best predict *interoceptive* signals – signals carrying information about the states of gut, viscera, hydration, vasomotor system, air-supply, muscular system, glucose and plasma levels, etc.

Here, the predictive processing (PP) account adds important dimensions to the well-known James-Lange model of emotional states as arising from the perception of our own bodily responses to external stimuli and events. The idea there, in a nutshell, was that our emotional 'feelings' are nothing but the perceptions of our own varying physiological responses. According to James it is our interoceptive perception of the bodily changes characteristic of fear (sweating, trembling etc.) that constitutes the very feeling of fear, giving it its distinctive psychological flavor. From a subjective viewpoint, interoceptive awareness manifests as a differentiated array of feelings including those of 'pain, temperature, itch, sensual touch, muscular and visceral sensations . . . hunger, thirst, and "air hunger"' (Craig, 2003, p. 500). The feeling of fear, if James is right, is thus essentially the detection of an interoceptive physiological signature that has already been induced by exposure to the threatening situation.

A popular (and useful) way to think about James' proposal is to see it as suggesting a kind of 'subtraction test'. This is a thought experiment in which you are invited to subtract all the bodily stuff (detection of your own racing heart etc.) away from the emotional experience, and ask yourself 'what would be left?'. James' claim is that you would be left with nothing that is worth counting as an experience or emotion. What an emotion really *is*, James argument suggests, is the *self-perception of changes in our own bodily states.*

But the standard Jamesian story remains somewhat inadequate. For it seems to require a one-to-one mapping between distinct emotional states and distinctive 'brute-physiological' signatures, and it seems to suggest that whenever the physiological state is induced and detected, the same emotional feeling should arise. Neither of these implications (see Critchley, 2005) has been borne out by observation and experiment. The basic story can, however, be refined and extended by adding a 'predictive twist'. Thus Seth (2013) suggests that a neglected core component may be the match (or mismatch) between a cascading series of top-down predictions of our own interoceptive states, and the forward-flowing information contained in sensory prediction error. Our interoceptive predictions, this story suggests:

> "arise from multiple hierarchical levels, with higher levels integrating interoceptive, proprioceptive, and exteroceptive cues in formulating descending predictions." (Seth (2013) p.567.)

A single inferential process here integrates all these sources of information, generating a context-reflecting amalgam that is experienced as emotion. Felt emotions thus integrate basic information (e.g. about bodily arousal) with higher-level predictions of probable causes and preparations for possible actions. In this way:

> "The close interplay between interoceptive and exteroceptive inference implies that emotional responses are inevitably shaped by cognitive and exteroceptive context, and that perceptual scenes that evoke interoceptive predictions will always be affectively coloured." (Seth, 2013 p. 563)

Physiologically, the Anterior Insular Cortex is remarkably well-positioned to play a major role in such a process by encoding what Craig (2003, p. 500) describes as 'a meta-representation of the primary interoceptive activity'. Emotion and subjective feeling states arise, this story suggests, as the result of multilevel inferences that combine sensory (interoceptive, proprioceptive, and exteroceptive) signals with top-down predictions to generate a sense of how things are for us and of what we might be about to do. Such a sense of 'action-ready being' encompasses our background physiological condition, estimations of current potentials for action, and the perceived state of the wider world. This delivers a grip upon both the nature and the significance our own embodied state.

Importantly, such a grip must integrate basic information (e.g. about bodily arousal) with higher-level predictions of probable causes. This provides a very natural way of accommodating large bodies of experimental results showing that the character of our emotional experience depends both on the interoception of brute bodily signals and higher-level 'cognitive appraisals' (see Schacter and Singer (1962), Prinz (2004)). An example of a brute bodily signal is generic arousal as induced by – to take the classic example from Schacter and Singer – an injection of adrenaline. Such brute signals combine with contextually-induced 'cognitive appraisals' leading us to interpret the very same bodily 'evidence' as either elation, anger, or lust according to our framing expectations.

## 2. Emotions as "constructs" (models)

The account of emotion just sketched fits perfectly with the *theory of constructed emotion* (Barrett, 2017). This mechanises Barrett's preceding *conceptual act theory* (Barrett, 2014) within a PP framework. The central claim is that in each waking moment the brain is integrating past experience to generate concepts to guide actions and give meaning to sensations. When the generated concepts involved relate to physiological imperatives, your brain constructs instances of emotion.

Following from the accounts of emotion in the PP literature, each instance of an emotion arises as a categorisation of bodily signals, according to context, in terms of past experiences:

> "When past experiences of emotion (e.g. happiness) are used to categorize the predicted sensory array and guide action, then one experiences or perceives that emotion (happiness)." (Barrett, 2017, p.9)

The theory of constructed emotion makes a sharp distinction between emotion *instances,* and emotion *categories*. An emotion *instance* is the in-the-moment construction of an emotion given the current context. What we usually describe as an emotion, (e.g. fear) is better described as an emotion 'category', which unifies diverse and highly variable instances under a single classificatory umbrella (Clark-Polner, Johnson & Barrett, 2016). Emotion categories, Barrett argues, do not exist in nature – they are assigned according to functional and socially constructed roles. Motivation for this view comes from what has been dubbed the "emotion paradox" (Barrett, 2006). The emotion paradox refers to the fact that while the existence of emotions such as "sadness", "anger", "happiness" is assumed by the scientific community and supported by common sense, the empirical literature calls into question this assumption due to the absence of any signature – be it a facial expression, physiological response or neural activity - that reliably indexes *any* emotion category. This leads to Barrett's claim that emotion categories are collections of diverse instances that are clumped together in terms of their functional role, lacking dedicated facial expressions, physiological responses or neural signatures, Barrett states:

> "Emotion categories are as real as any other conceptual categories that require a human perceiver for existence, such as 'money' (i.e. the various objects that have served as currency throughout human history share no physical similarities)." (Barrett, 2017, p.13).

This many-to-one mapping of physical states to emotion categories - called 'degeneracy' - is the primary argument behind the lack of any kind of emotional "essence". Degeneracy is borne out by the empirical literature. A meta-analysis of facial expressions indicates that many different facial expressions can be observed for the same category, and many different emotional categories can be understood by the same facial expression (Duran et al, 2017) – the meaning of a facial expression largely depends on context. Physiological signatures for any emotion category have proved to be similarly elusive, with a recent meta-analysis (Siegel et al, 2018) showing that there are no physiological signatures that reliably correspond to any one emotion category – for instance, when you're angry, your blood pressure can go up, down, or remain the same. On Barrett's view the determining factor is what kind of *action* the brain is preparing the body for – getting ready to fight requires recruitment of different resources than some other anger-related course of action, despite the emotion categorisation ('anger') being the same (Barrett, 2017). Similarly, a meta-analysis on the neurophysiological basis of emotion categories are not contained within any one brain region or system, but are represented as configurations across multiple brain networks (Wager et al, 2015).

From the perspective of evolution, degeneracy in the brain makes sense as an adaptive engineering principle. A key result of degeneracy is that a single brain can create a vast number of spatiotemporal patterns. These high complexity systems are preferred by natural selection as they can as they can reconfigure themselves into a multitude of different states (Whitacre, 2010; Whitacre and Bender, 2010). This reconfiguration ability is what makes our brains, on this account, radically flexible according to culture and environment.

Emotions, then, are not reactions to the world, not even *interoceptively informed* reactions to the world. Rather, they are out-and-out constructions of the world. Emotions are constructed in just the same way that percepts are constructed; that is, they are predictive models of the likely causes of the sensory input, made by re-stitching together past experiences and then classifying the current experience as an amalgam of past experiences of a similar nature. These emotional predictions are made always in the service of regulating the body's internal milieu, that is, in the service of *allostasis (*Barrett and Simmons, 2015; Barrett, 2017) Predictive processing, Barrett suggests, provides the mechanism underlying these categorisations.

On this more 'action-oriented' predictive processing account, the top-down flow of predictions anticipate 1) upcoming interoceptive and exteroceptive signals and 2) the best action or bodily response to deal with the upcoming sensory flow. In order to create these 'concepts' (embodied, whole-brain representations), the brain creates predictions by using past experience to answer *"What is this new sensory input most similar to?"* (Barrett, 2017). The incoming sensory evidence, in the form of prediction error, helps to select and shape the distributions of predictions that are activated that best fit the sensory array, thereby minimising prediction error – resulting in a *categorisation* of the incoming sensory information in terms of past experiences (Barrett, 2006). That means that the predictions activated in the present are an instance of what Barsalou refers to as 'ad hoc' concepts (Barsalou, 1983). In the brain, a concept looks like a distributed pattern of activity across populations. These ad hoc concepts or predictions, that categorize present sensory flux in terms of past experience, are the mechanism of construction of any given instance of emotion. This predictive cascade – the interpretation of the sensory flux in terms of its expected utility to allostasis - is the process of meaning-making in the brain.

Notice also that emotion and cognition are here performed in exactly the same way, that is, in reference to allostasis, and sensory inputs (prediction error) are used as information to guide the sculpting of concepts that engender adaptive action. This process is an approximation of Bayesian inference (Deneve, 2008) to decide amongst which simulation (interlocked web of predictions) should be implemented in order to maximise allostatic efficiency across multiple body systems (e.g. need for glucose, oxygen, salt etc.), and activate appropriate metabolic expenditure in the service of action (tiger, run!).

Barrett's theory is supplemented with a compelling neurobiological implementation story, where the default mode network represents efficient, multimodal summaries, which, when activated, cascade through the entire cortical sheet, terminating in primary sensory and motor regions. The cascade as a whole is an instance of a concept, or an emotion (Barrett, 2017). That said, the link between the neurobiology and the conceptual argument is not altogether clear: the empirical evidence is open to interpretation and amenable to other conceptual theories of emotion (including other conceptual theories with PP as the underlying mechanism).

The theory of constructed emotion offers a plausible account of how diverse instances of emotion come to be placed together under unifying conceptual umbrellas. It also fleshes out how emotion categories are cleaved apart according to context, and how the categories are more socially determined conceptual categories than categories existing in nature. Furthermore, the theory partially fleshes out the conception of emotion as interoceptive inference, both with a more specific mechanism of diverse instances of emotion, and in setting out how different emotion categories come to be formed.

## 3. From Embodied Emotion to Embodied Valence

So how do we make sense of affective value or valence? What determines the evaluative dimension of an emotion instance? Here is an initial approach we might take to accounting for valence in terms of the properties of an action-oriented predictive processing system.

The core imperative of a predictive processor is the successful prediction of incoming sensory evidence. Thus it may initially seem that the successful minimisation of prediction error should be what determines an overall state of positive valence. Though this may seem promising at first, such a

proposal quickly falls apart. Any account of valence that is state-based, that equates positive valence to a state of minimized prediction error, fails to do justice to the fact that prediction error minimisation is necessarily a dynamic and continuous process, constantly engaging action, and designed to account for the on-going maintenance of an organism in an ever changing world. Only from this perspective can we avoid the 'dark room' objection to predictive processing (Friston, Thornton, and Clark, 2012). This states that if my goal is solely the minimisation of prediction error, then surely I should just seek out a dark, empty room and stay there. Perfect prediction, it seems, is attainable by avoiding action and practicing sensory (and nutritional) deprivation until death. Such a policy is, of course, wholly inconsistent with the actual behaviour of living things.
An initial response to this might be that the various demands of survival (as ultimately signalled in the form of prediction error) would move you onwards. But note that even were your dark room to come equipped with a life support machine (consider an unending night in an abandoned hospital ward) it is unlikely that you would find this to be an endlessly pleasurable experience. Humans not only find a lack of novel stimulation boring, they actively seek out and take delight in a rich repertoire of aesthetic, humorous, or thrilling situations, from skydiving to stand-up, that are specifically engineered to generate a rush of prediction error through the violation of prior expectations.

A more promising strategy is as follows. Instead of tying valence to the achievement of some particular error-minimized state, Joffily and Coricelli (2013) propose a dynamic alternative in which valence is taken to be the *rate* at which this error is being reduced. In mathematical terms valence is recast as the first time-derivative of error: a matter of *velocity*, rather than position. We seek out surprising states, then, in as much as they offer us the opportunity to engage in a faster (rather than slower) rate of reduction in prediction error. Drawing on Carver and Scheier's (1990) control theoretic account of emotion, Van De Cruys (2017) improves and extends this story by suggesting that, rather than being straightforwardly a matter of a positive rate of error reduction, pleasure (positive valence) occurs when our actual rate of error reduction is higher than we had predicted it would be. If it is lower, we experience negative valence.

An upshot of explaining valence in terms of these processing characteristics, rather than specific content, is that it is no longer tied to any particular set of causes, error modality, or level of inference. We can thus describe a relationship between allostasis and valence that is not constrained (as it is in Seth (2013)) to inference over the causes of interoceptive signals alone. This seems like the correct route to take. Homeostatic maintenance is served not only by the direct monitoring and regulation of physiological variables, but also indirectly, by the anticipatory regulation of our external environment. Whether intero or extero-ceptive, persistent unreduced prediction error is a sign that we do not have a grip on our self or surroundings, and adjustments need to be made. Furthermore, tying valence to the regulation of exteroceptive error reduction rate allows us to characterise more 'cognitive' experiences of positive valence – those that are not easily described in terms of basic physiological reactions—such as responses to art, narrative or humour. These can now be understood as achieving their emotional effects by engineering pleasurable trajectories in the creation and violation of expectations, followed by the subsequent pleasurable release in the eventual reduction of resulting prediction error. This fits nicely with descriptions of humour, as resulting from the creation and resolution of tension (Sroufe and Waters, 1976) and, as Van De Cruys and Wagemans (2011) suggest, provides a potential explanation of the failure of aesthetic principles (such as harmony, fluency, or balance) to account for the success of celebrated works of art which regularly display the deliberate violation of such rules.

## 4. Emotion and Cognition

Summing up the previous sections, what predictive processing reveals is a world permeated by affect – a world of opportunities for action, geared to current tasks, modulated by information about our own bodily states. But to see just how radical the PP picture turns out to be, we still need to add one final ingredient. It's that PP rejects the picture of emotion and cognition as fundamentally different kinds, at least insofar as they are causally active parts of the cognitive machinery.

According to a popular view, often associated with Hume, a fundamental divide among all things mental is one that divides the informational and the motivational. The former is about the organism ("coldly") coming to a view about what's going on in the world, whereas the latter is about ("hotly") driving the organism to bring about change in the world. Hume's central point was that without the latter, without passions, no action would ever take place. A hypothetical creature only capable of having informational states would stay still, inert and unmoved to do anything, regardless of what it learnt about the world. In this sense, according to Hume, emotion (affect, passion) broadly construed, is the driving force behind all action, but completely distinct from belief (and insulated from "reason").

The idea that informational states on the one hand, and motivational states on the other, are fundamentally different kinds of state whose interaction is required to bring about action, is widely embraced in daily life. It forms not only a core part of common-sense (or 'folk') psychology, but is deeply embedded in some more scientific frameworks too. Statistical decision-theory (including neuroeconomics and work on reinforcement learning) inherits this Humean picture, since in standard realizations it works with a firm separation between encodings of value or 'utility' and encodings of probability. In these frameworks, decisions are made and actions selected only when utility and probability align, revealing viable opportunities for worldly interventions that deliver weighted rewards at calculated costs (for a useful review, see Sanfey et al (2006)).
By contrast, PP posits only predictions, informed by multiple inner and outer sources of information. In PP motivational states are realized as predictions about our own future actions and states. To see how, let's return to the PP treatment of action. Action is making the world conform to some of your predictions, and is just another way of reducing long-term prediction error. At the bottom level, PP makes sensori-muscular (proprioceptive) prediction into a proxy for motor commands (Shipp, Adams and Friston (2013)). Predicting the flow of sensori-muscular effects that would occur if you hit the tennis ball just right actually brings the 'good hitting' about. In a little more detail, the brain predicts the flow of states of muscle spindles, tendons, and joints that the action demands, and the resulting errors (since those states are not yet actual) are systematically quashed by moving the body so as to make that flow of prediction come true. This is an elegant and economical means of delivering basic motor control (see e.g. Shipp et al. (2013)).

PP deploys the same kind of story 'all the way up'. Our action-guiding proprioceptive predictions are themselves caused by even higher-level and longer time-scale predictions – predictions about our own future behaviors and future states. These entrain actions when good opportunities arise (see Pezzulo, Rigoli and Friston (2015)). The picture is of nested beliefs that entrain actions by bringing about predicted sensory flows. For example, suppose I believe/predict that I will meet you at 7:00 at the movie-theatre. This (combined with prior knowledge and any newly gleaned information) leads me to believe/predict that I will get the 6:30 bus. That last prediction then acts as a kind of minipolicy that enslaves motor action (by means of proprioceptive predictions) when it is time to leave the house.

Simple action-entraining motor intentions here cash out as precise proprioceptive predictions, while higher-level intentions, including standing goals, are realized by higher-level predictions of whole swathes of sensory information, which likewise entrain actions (by yielding precise proprioceptive predictions) when they themselves are assigned high enough precision. These nested, interacting predictions arise and dissolve, in ways that realize the phenomenological flux of shifting drives and desires, as we move around the world, acting and harvesting new sensory information. If PP is on track, the causally potent play of human motivation is not an illusion—but it is realized using only the common currency of multi-level, multi-area prediction. In this picture, prior beliefs (resulting in predictions) combine with sensory evidence to bring about action. This is just the bedrock (Bayesian) move – one that turns everything into a form of prediction-based inference.

It has been suggested (Holton (2016), Klein (2016)) that this picture is too impoverished to be a satisfying story about human minds. Part of their reasoning is roughly Humean. The Humean worry is that beliefs (or predictions) without motivations are inert, unable to mandate actions. That's already taken care of by the PP story though, since high-precision predictions that have proprioceptive (hence motoric) consequences are immediately poised to entrain actions to make themselves come true. Holton also worries that assimilating desires to predictions "doesn't do justice to the multiplicity and malleability of human desire" noting that we need to accommodate cases where we desire X and may even do X while believing that X won't bring us happiness or pleasure. However, PP accommodates this very simply, by separating predictions about the hedonic consequences of actions from the full set of predictions that interactively entrain actions. Specifically, the predictive processing story firmly distinguishes (Friston, Shiner et al. (2012)) between sub-personal action-entraining high-precision predictions concerning what I will do and predictions of the hedonic (interoceptive) outcomes of those very actions. PP thus accommodates the fact, highlighted by Holton, that drug users often do not believe/predict that taking the drugs will actually lead to happiness. But what they do predict is seeking and ingesting the drug. PP thus easily reconstructs the useful distinction between 'wanting' and 'liking' (Berridge (2007)). The PP picture thus turns out to be a neat fit with important work on the nature and mechanisms of addiction (Berridge (2007), Friston et al (2012)). More generally, even given that the addict need not predict that the drugs will bring pleasure, PP remains poised to explore a wide variety of promising accounts in which the addict's experiences and actions are the results of interacting sub-personal (non-conscious) predictions.

This replaces Hume's two interacting kinds (reason and passion) with a picture of large numbers of subtly different and modifiably interacting elements. All of those elements are somewhat belief-like (consisting in predictions) but somewhat desire-like too (as they help select and entrain actions at multiple time-scales). So, while it may look like a simplifying move, what PP finally delivers will in fact be a far richer palette for explaining human behaviour. That palette allows a full spectrum of possibilities that reach far beyond the simple, constrained interactions suggested by crude folk psychological distinctions between 'cognition' and 'conation'.

We have seen how this collapses belief and desire, and desire is often construed as a "hot" or "impassioned" state, but it is clearly a mistake to equate emotion with desire. As several theorists have noted (e.g. Marks 1982, Oakley 1992), emotion has both belief-like and desire-like elements. Experiencing fear of the spider simultaneously tells you about the world (e.g. that there is a spider there), while also motivating you to act in a certain way (run away from said spider). But whereas the standard way of thinking of emotions is as *composed of* these belief and desire-like elements, PP construed things very differently. Just because the belief and desire-like elements can be "read off" the emotional state, it is not to say that psychologically (or indeed ontologically) they are somehow the primitive building blocks of a hybrid and less primitive state called emotion. On the contrary, according to PP, it is the emotional state, which simultaneously informs and moves, that is primitive, and, in predictive processing terms, this is all fleshed out in the common currency of predictions and predictive models: predictions generated by complex hierarchical models concerning, in an interconnected manner, the organism, the world, and the organism's place in that world.
The same point can be made in terms of direction of fit. Whereas it has been common to think of beliefs, with their mind-to-world (or descriptive) direction of fit, and desires, with their world-to-mind (or directive) direction of fit, as being the fundamental building blocks of the mind, what is actually fundamental in the PP architecture is prediction, which can vary across a spectrum as to the extent to which it should be fulfilled by the world (perception/belief) or the self (action/desire). This means that pure belief (or cold perception), or pure desire (or blind action), is actually

idealization, a mere theoretical construct. What we are actually left with is a wide variety of what Millikan (1995) calls "pushmi-pullyu representations", states that simultaneously describe and direct.

## 5. Conclusion

Emotions, we have argued, are built from predictions. They reflect inner and outer sources of information, combined in flexible ways, and are answerable to the full world knowledge (generative model) of an agent. But they are not a special cognitive kind. Instead, they are part and parcel of an integrated processing system whose core functionality is to reduce precision-weighted prediction error by maintaining dynamic engagements with the world. These engagements display trajectories both marked and determined by valence, where positive valance reflects better-than-predicted slopes of error minimization. What emerges is a picture of mind as an action-oriented system all of whose states are somewhat belief-like, and somewhat desire-like too.

Another way of looking at this is as follows. In so far as full-blown emotions as we typically understand them are the most prominent and consciously detectable (and hence categorized) of these action-oriented states, one could say that PP renders emotion, construed more broadly to include even the very subtlest of these, ever-present in cognition. In other words, the embodied predictive mind is, by necessity, an emotional mind.

# Chapter 3: Dissolving the self: Active inference, psychedelics, and ego-dissolution

Psychedelic drugs such as psilocybin, LSD and DMT are known to induce powerful alterations in phenomenology. Perhaps of most philosophical and scientific interest is their capacity to disrupt and even "dissolve" one of the most primary features of normal experience: that of being a self. Such "peak" or "mystical" experiences are of increasing interest for their potentially transformative therapeutic value. While empirical research is underway, a theoretical conception of the mechanisms underpinning these experiences remains elusive. In the following chapter, psychedelic-induced ego-dissolution is accounted for within an active inference framework, as a collapse in the "temporal thickness" of an agent's deep temporal model, as a result of lowered precision on high-level priors. The argument here is composed of three moves: first, a view of the self-model is proposed as arising within a temporally deep generative model of an embodied organism navigating an affordance landscape in the service of allostasis. Next, a view of the action of psychedelics as lowering the precision of high-level priors within the generative model is unpacked in terms of a high Bayesian learning rate. Finally, the relaxation of high-level priors is argued to cause a "collapse" in the temporal thickness of the generative model, resulting in a collapse in the self-model and a loss of the ordinary sense of being a self. This account has implications for our understanding of ordinary self-consciousness and disruptions in self-consciousness present in psychosis, autism, depression, and dissociative disorders. The philosophical, theoretical and therapeutic implications of this account are touched upon.

## 1. Introduction

Psychedelic ("mind-manifesting") drugs are known to occasion radically altered states of consciousness, including profound changes in sensory perception, emotion, cognition, time perception, and self-consciousness (Preller & Vollenweider, 2016). One of the most interesting of all of these effects is the experience of ego-dissolution. Although the experience is notoriously difficult to articulate and even considered ineffable, psychedelic researcher Stanislas Grof, who considers ego-dissolution the “main objective” of psychedelic therapy, describes it as “an ecstatic state, characterized by the loss of boundaries between the subject and the objective world, with ensuing feelings of unity with other people, nature, the entire Universe, and God” (Grof, 1980, p. 79). Ego-dissolution is of considerable philosophical and theoretical value for understanding selfhood and the nature of consciousness (Letheby & Gerrans, 2017; Millière, 2017; Nour & Carhart-Harris, 2017). It is also considered to be central to the therapeutic potential of psychedelics (Carhart-Harris & Goodwin, 2017; see also Letheby, 2020, Limanowski & Friston, 2020, Sebastián, 2020, all in this special issue). Despite this, very little is known about the mechanisms underpinning psychedelic-induced ego-dissolution. “Predictive processing” theories of brain function (Clark, 2013; Friston, 2010; Wiese & Metzinger, 2017) have recently taken precedence in cognitive science, affording a novel theoretical framework to approach cognitive phenomena. In this chapter I propose a novel account of ego-dissolution within an active inference framework. To this end, I initially furnish an account of self-modelling within active inference, where pre-reflective self-consciousness emerges in organisms as a consequence of “temporal thickness”, the need to model the consequences of potential actions over time (Friston, 2018). I then give an account of the action of psychedelics within a predictive processing framework, unpacking the view that psychedelics “relax” high-level priors (Carhart-Harris, 2019; Carhart-Harris & Friston, 2019) in terms of a high Bayesian learning rate (Hohwy, 2017; Mathys et al., 2014). Finally, I argue that low precision at high-levels of the inferential hierarchy results in a collapse of the temporal thickness of the generative model and the corresponding self-model, leading to the phenomenon known as ego-dissolution (see also Limanowski & Friston, 2020)

## 2. The free energy principle

The Free Energy Principle (FEP) has the most ambitious explanatory scope of all “predictive processing” style theoretical frameworks (Friston, 2010). It combines, subsumes and links to several brain theories, including the Bayesian brain hypothesis (Knill & Pouget, 2004), predictive coding (Mumford, 1992; Rao & Ballard, 1999), efficient coding (Barlow, 2001) and reinforcement learning (Dayan & Daw, 2008). The mathematics of the theory are complex and beyond the scope of this chapter (for a review see Buckley, Kim, McGregor, & Seth, 2017). According to the FEP, simply in virtue of existing, all organisms tend to minimise the entropy or dispersion of their states. This much is intuitive: the conditions that are viable for an organism are fairly narrow – deviation from homeostatic bounds, such as having a body temperature of 50 degrees centigrade, is incompatible with continued existence. Organisms that fail to stay within their "species-specific window of viability" (Clark, 2013 p. 13) simply cease to exist. Life, on this account, resists the tendency towards disorder imposed by the second law of thermodynamics, and this principle applies at all levels – "from their gross morphology to fine details of cortical microcircuitry as well as at timescales from the neuronal to the phylogenetic" (Seth & Tsakiris, 2018, p. 973). Organisms, then, must resist entropy, the long-term average of (information-theoretic) surprise. Because this quantity is beyond direct epistemic access to an organism, according to the FEP, organisms minimise a proxy variable or upper bound – dubbed (variational) free energy. Free energy (under some simplifying assumptions) is equivalent to precision-weighted prediction error in predictive processing.

On the predictive processing view, the brain has stored prior beliefs (in the form ofprobability distributions) about the causes of sensory inputs in the world (Clark, 2013; Wiese & Metzinger, 2017). Prior beliefs are hierarchically organised, where higher-levels encode predictions about representations at lower levels. Prediction errors, arising from the discrepancy between the low-level predictions and incoming sensory signals, are passed up the hierarchy, where higher-level predictions are updated to minimise further prediction errors. Perception, then, both exteroceptive and interoceptive, is the product of (approximate) Bayesian inference, whereby the influence of prior beliefs and sensory evidence are weighted according to "expected precision", e.g. confidence in the given context, to generate a posterior. Inference in these schemes is thought to occur across a hierarchy of inferred causes, where higher levels encode regularities that occur at larger spatial and temporal scales (Kiebel, Daunizeau, & Friston, 2008). In perceptual inference, sensory prediction errors can be minimised by tweaking the parameters of the generative model – that is, generating predictions to quash the influx of prediction error. Prediction error can also be minimised though action by changing the incoming sensory data to fit a prediction – for instance, I can move my eyes to bring my coffee cup into view, to fulfil the prediction of a coffee cup. Actions can be thought of as the fulfilment of proprioceptive (or oculomotor) predictions – an intended movement occurs as a result of predicting the proprioceptive consequences (Friston et al., 2010; Shadmehr, Smith, & Krakauer, 2010). There are detailed accounts of the neural implementation of these schemes available (Bastos et al., 2012; Keller & Mrsic-Flogel, 2018; Shipp, 2016).

### 2.2 Precision-weighting

A key feature of predictive processing schemes is the contextual flexibility afforded by precision-weighting (Clark, 2013; Feldman & Friston, 2010). Precision regulates the interaction between top-down and bottom-up signals, through the synaptic gain on neuronal populations signalling prediction error, in order to approximate optimal inference over time. Precision can be thought of as tracking both the reliability and relevance of the incoming sensory information, where weighting by reliability is analogous to assigning greater weight to more reliable information when updating a belief. Prediction error signals with high precision (inverse variance) have greater influence in updating the top-down predictions. Precision itself has to be inferred, both by the empirical variance in the sensory data itself, and according to prior expectations about precision. The optimisation of precision weighting, through updating of the precision expectations (precision-related priors), is frequently equated to attention within predictive processing (Clark, 2013; Feldman & Friston, 2010). Importantly for the current treatment, precision is thought to mediate both sensory attenuation—the top-down filtering out of afferent information, and affordances, where affordances refers to the latent possibilities for action given the capabilities of the agent (Cisek, 2007; Friston, Shiner, et al., 2012).

### 2.3 Control-oriented inference

In mandating that existence necessitates maintaining oneself within a limited repertoire of states via control-oriented predictive regulation (instrumental active inference) (Seth & Tsakiris, 2018), the FEP aligns itself with precursors of this view, cybernetic theories that build on control, feedback and predictive modelling (e.g., the "good regulator theorem") (Conant & Ashby, 1970). Note that while a purely Helmholtzian view of the brain might cast it in terms of inferring hidden causes in the world, casting the predictive machinery in terms of being for ensuring continued existence means that the generative model is not constrained to veridicality. Rather than faithfully reconstructing the world, perception is "ultimately geared towards driving actions that preserve [the] physiological integrity of the organism. In other words, we do not perceive the world (and self) as it is, but as it is useful to do so" (Seth & Tsakiris, 2018, p. 975).

### 2.4 Homeostasis

Homeostasis refers to the tendency of living systems to keep an "internal balance" despite changes in the surrounding environment (Cannon, 1929). This has long been described in terms ofcontrol theoretic and cybernetic mechanisms, and more recently this homeostatic control is thought to involve interoceptive signals that report current physiological states (e.g., heart rate, or blood-bound glucose levels) (Craig, 2002; Damasio & Carvalho, 2013). One way to restore bodily conditions to favourable states is to engage autonomic reflexes – for example, a hyperthermic animal can perspire to cool down. Of course, autonomic regulation alone is not sufficient to ensure continued existence – to avoid hunger or thirst the animal must engage actions, such as seeking out food and water. Collectively, these actions are termed allostasis, the process via which the brain regulates the needs of the body (Corcoran & Hohwy, 2018; Corcoran, Pezzulo, & Hohwy, 2020; Schulkin & Sterling, 2019). Crucially, to stay viable on longer timescales, this action must be anticipatory – avoiding dyshomeostatic conditions before they arise (Pezzulo, Rigoli, & Friston, 2015; Sterling, 2012).

### 2.5 Active inference

The FEP regards homeostasis and allostasis as the central aspects of organic life, thus the autopoietic principles at the basis of the FEP act as a kind of "first prior" (Allen & Tsakiris, 2018). In other words, "[t]he brain is in the game of predicting the world, but only as a means to the end of embodied self-preservation" (Allen & Friston, 2018, p. 12). In so doing, the free energy principle collapses expected utility (instrumental value) and information gain (epistemic value) under a single quantity. On this approach, action planning is itself a form of inference, where preferences and goals are framed in terms ofprior beliefs, such that these priors are fulfilled by action (Botvinick & Toussaint, 2012). Casting value and utility purely as inferential problems may at first appear unintuitive – if an agent finds itself in consistently adverse circumstances, then such adverse circumstances should, at first pass, seem to have high probability. However, "[t]he critical step in this logic is the assumption that evolution has equipped us with the belief that low utility states are low probability, due to the fact that if our ancestors spent a lot of time in those states they would be less likely to reproduce" (Gershman, 2019, p. 7). The so-called "first prior", that of maintaining existence via homeostatic and allostatic regulatory behaviour, ensures that organisms seek to actively maintain internal and external conditions conducive to their own persistence.

Active inference refers to the process by which agents actively sample states of the world so as to reduce uncertainty and realise prior preferences, rendering the action selection process itself an inference problem. This arbitration occurs according to priors pertaining to expected free energy over a given course of action, or policy (Friston et al., 2018; Pezzulo et al., 2015). Expected free energy is the free energy an agent predicts itself to average in opting to pursue a particular course of action. Intuitively, some courses of action are more likely than others to lead to "expected" or desirable outcomes. A policy that has lower expected free energy is going to have a higher prior probability than a policy with higher expected free energy, because agents equipped with prior beliefs about their continued existence will pursue policies that reduce expected free energy (Friston et al., 2015, 2018; Kaplan & Friston, 2018). Crucially, agents engaging in active inference do not merely restrict themselves to the states they expect; rather they anticipate in order to minimise uncertainty about potential future outcomes (Friston et al., 2015, 2018; Pezzulo et al., 2016). This prospective form of control relies on the contextualization provided by higher levels in the inferential hierarchy, which anticipate the downstream consequences of actions and select policies accordingly (Friston, 2010; Pezzulo et al., 2015). Contextualisation here depends on the relative precision at various hierarchical levels, where "precision dynamics subsume the role of arbitration" (Pezzulo et al., 2015, p. 27). This approach bears similarities to other control-theoretic approaches, such as the affordance competition hypothesis (Cisek, 2007; Pezzulo & Cisek, 2016), where an affordance is a potential for action that avails itself to an organism in its action-oriented perception of environmental features.3 On this view, perceived affordances jostle for precedence and are arbitrated on the basis of the desirability of their predicted outcomes.

## 3. The self in active inference

This section outlines an account of how pre-reflective self-consciousness – an implicit sense of being a subject present in all experience – is structured within an active inference framework. Here, the self-model is underpinned by the same inferential Bayesian schemes that are increasingly being used to describe perception and action. This predictive-modelling approach to selfhood has roots in Thomas Metzinger's work on conscious and unconscious self-models, and the "self-model theory of subjectivity" (Blanke & Metzinger, 2009; Metzinger, 2003, 2009) where "[a] self-model, an inner image of the organism as a whole [is] built into the world-model, and this is how the consciously experienced first-person perspective develop[s]" (Metzinger, 2009, p. 64). The account presented here follows the increasing focus on the embodied nature of selfhood, where "being" or "having" a body is thought to be one of the most basic aspects of the experience of being a self (Allen & Friston, 2018; Apps & Tsakiris, 2014; Blanke & Metzinger, 2009; Limanowski & Blankenburg, 2013). A growing number of researchers seek to ground selfhood and emotion in interoceptive processes, particularly in their functional relation to allostatic regulation (Barrett & Simmons, 2015; Seth, 2015; Seth & Friston, 2016). A key reason for this is that interoceptive inference is apt to put greater emphasis on control over discovery (Seth & Friston, 2016), due to "a priori hyper-precision of visceral channels"(Allen & Friston, 2018, p. 7), in which interoceptive signals are assigned very high precision in virtue of communicating information about key physiological variables (Seth, 2015). Grounding the self-model in control-oriented active inference (Seth & Tsakiris, 2018) inflects perception of the affordance landscape in terms of bodily states, an idea which is nicely expressed by Montague and King-Casas:

> A sated and comfortable lioness looking at two antelopes sees two unthreatening creatures against the normal backdrop of the temperate savanna. The same lioness, when hungry, sees only one thing – the most immediate prey. In another circumstance, in which the lioness may be inordinately hot, the distant, shaded tree becomes the prominent visual object in the field of view. (Montague & King-Casas, 2007, p. 519)

This forms the basis for the view that will be unpacked in more detail in what follows, that the self-model can be understood as an "allostatic control model", arising from the system's sense of control of the temporally deep consequences of actions for allostasis. On this view, pre-reflective self-consciousness is underpinned by the inference about endogenous control of the sensory consequences of actions within deep goal hierarchies, where goals and preferences are framed in terms of prior beliefs, such that goals are fulfilled by actions (Botvinick & Toussaint, 2012; Pezzulo et al., 2015). Recall, action allows an organism to change the sensory input in order to conform to its generative model, as opposed to perceptual inference that involves revising model parameters to conform to the sensory input. In order to act, then, the system implicitly infers its own ability to bring about the intended sensory consequences – it is in this sense that "implicit in a model of sampling is a representation or sense of agency" (Friston, 2012a, p. 173), which is closely related to what has been called the "primacy of the 'I can' " (Bruineberg, 2017). Crucially, organisms with deep temporal models have "temporal thickness" – expectations regarding the sensory consequences of actions on multiple interlocking timescales. The following sections will unpack this conception of the self-model in terms of hierarchically deep allostatic control, starting with the notion of temporal thickness, and then moving to how motivated control hierarchies "attune" organisms to action opportunities on multiple timescales, for both proximal goals, for instance, pain motivating an organism to act so as to fulfil a "healthy body condition" prior; and distal goals, for example emotions motivating a change of circumstances pertaining to longer timescales such as moving to a different city. The discussion will then move to how deep self-models allow organisms to arbitrate between different policies and trade off outcomes on different timescales.

### 3.1 Temporal thickness

To successfully navigate the world over longer timescales, and select policies that result in survival – and not dispersion or non-existence – organisms must possess models of the future; in other words, they require deep temporal models (Friston et al., 2018). The generative models that endow organisms with the capability of inferring the consequences of future actions must have the property of temporal thickness, which allows the organism to anticipate the downstream consequences of potential actions, conferring the ability to select policies or action scripts that are favourable to the organism's continued existence (Friston, 2018). The minimisation of surprise through active inference on the FEP involves acting so as to reduce uncertainty, and to do this the system must model itself across time and counterfactuals – that is, it must model what kind of agent it is at varying degrees of temporal depth. Self-modelling, then, emerges as a natural consequence of prospective action selection (Friston, 2018), where the principal function of a counterfactually rich self-model is to facilitate navigation of the affordance landscape and action selection across multiple interlocking timescales – for example expectations of what an agent can do on shorter timescales inform expectations of what the agent can do over longer timescales. The functional role of having a rich self-model, then, is that it enables the organism to predict outcomes across diverse policies, and endows the organism with "what if?" capabilities (Friston, 2018), which puts this picture into contact with mental time travel and offline simulation (Buckner & Carroll, 2007; Schacter, Addis, & Buckner, 2008).

### 3.2 Attuning to the world

Conceiving of the self-model through an active inference framework, a hierarchically deep self-model guides policy selection over various timescales in service of minimising expected free energy. In what follows, pain perception, viewed as arising through the violation of the prior of "healthy body condition" (Ongaro & Kaptchuk, 2019), will be used to illustrate how inferences about the self "attune" an organism to adaptive action opportunities. One key advantage that the active inference account of self-modelling has over strictly Bayesian approaches is that it is goal-directed (Moutoussis et al., 2014). Classical models of pain perception as the consequence of physiological dysfunction are challenged by the efficacy of placebo treatments in relieving pain (Anchisi & Zanon, 2015), and cases in which pain is experienced without physiological disruption, as is often the case in chronic pain. Instead, there is evidence to suggest that affectively charged percepts, such as pain, are best understood as resting on the same inferential mechanisms as are assumed to underpin perception and action under a predictive processing framework (Büchel, Geuter, Sprenger, & Eippert, 2014). Bayesian models of pain perception (Morton, El-Deredy, Watson, & Jones, 2010) indicate that prior beliefs about the generation of painful percepts are integrated with current sensory data to infer the posterior or hidden worldly cause (the painful percept). Crucially, these pain percepts incorporate the "weight" or precision of past experiences when computing the current painful percept (Morton et al., 2010). On their own, however, these models of pain perception are silent on the functional role of pain as a motivator to an embodied organism (Moutoussis et al., 2014). Optimal inference about pain to the allostatically concerned organism is heavily dependent on the context, as anyone who has felt the pain of an injury only after danger is averted can attest to. In this way, pain perception is allostatically "tuned": "organisms can tune their own pain perception according to both their prior beliefs and the specific biological goals they believe are attainable in that context" (Moutoussis et al., 2014, p. 70).

A Bayesian framework of pain perception, therefore, needs to represent the agency and aims of the organism. This is precisely what is afforded by conceiving of the self-model within an active inference framework (Friston, 2012b) – as this provides the necessary context to study the self-model, across multiple hierarchical levels. Like physical pain, and sharing the neural underpinnings of physical pain (Eisenberger, 2012), social pain is similarly understood in inferential terms, and does not scale with "damage" per se (for instance, social rejection), as evidenced by the wide range of sensitivity people have to the same physical manipulation (Eisenberger, Jarcho, Lieberman, & Naliboff, 2006). Accordingly, there is evidence to suggest that appropriately "tuning" emotional responses in social contexts allows for agents to approximate Bayesian inference in policy selection given bounded cognitive capacity and rationality. For example, on a "stag hunt" game,4 agents with "prosocial" preferences can outperform agents of similar cognitive sophistication that lack social biases (Yoshida, Dolan, & Friston, 2008).

### 3.3 Emotion

Conceptualising emotions in terms of a contextualisation of bodily states has historical roots dating back to the James-Lange theory ofemotion (Cannon, 1927) and two-factor theory of emotion (Schachter & Singer, 1963). Lisa Feldman-Barrett has developed this approach specifically within the active inference framework as the "theory of constructed emotion" (Barrett, 2017). According to the theory of constructed emotion, emotions are constructed in the same manner as percepts, where priors are recruited according to context to make a "best guess" at the hidden causes of (interoceptive) sensory signals. On Barrett's view, emotions arise through a context-sensitive inferential categorisation of interoceptive states. For this reason, emotions on this view are "constructions" – there are no neural or physiological signatures that reliably discriminate any emotional state (Siegel et al., 2018; Wager et al., 2015). Rather, physiological reactions in the body occur in order to prepare it for action, and these are categorised as emotions only contextually through the predictive models recruited to explain away the incoming afferent interoceptive signals. For example, heart rate increases or decreases depending only on an anticipated action – e.g., fight or flight – and given an emotional ascription only contextually – e.g., the same bodily state could be categorised as fear in one context and anger in another. Interoceptive inference is experienced as emotion in service of producing allostatic action (Barrett, Quigley, & Hamilton, 2016; Barrett & Simmons, 2015; Wilkinson, Deane, Nave, & Clark, 2019). In viewing the self-model in terms of hierarchical allostatic control, interoceptive inference on the hidden causes ofbodily states pertaining to longer timescales tunes perception to the world and affordances differently, such that more abstract emotions might track regularities over longer time scales, informing policy selection thereon (Pezzulo, 2014), and allowing for more abstract and distal outcomes to be motivationally salient.

### 3.4 Hierarchically deep self-models

Viewing the self-model in terms of allostatic control renders selfhood fundamentally affective and action-oriented, such that different aspects of the self in a given context – precision on goals and preferences at different levels of the hierarchy – motivate behaviour and arbitrate between policies. On this view, the self-model inflects perception of possible actions in the world and mediates salience to facilitate the selection of policies with minimal expected free energy. Critical to this picture is the notion that these various models are associated with varying degrees of temporal depth (Pezzulo, Rigoli, & Friston, 2018). Deep generative models capture increasingly distal relations between actions and outcomes within hierarchical active inference, allowing for the coordination of behaviour across different hierarchical levels, enabling goals to become prioritised relative to current context (Pezzulo et al., 2018). The result is an inferential framework of hierarchically nested contextual complexity, in which lower levels track basic (and sometimes evolutionarily hard-wired) motivations or affordances, while higher levels track motivations and plans over deeper timescales. In this way, higher-level contextualization of lower sensorimotor functions optimises expected actions in terms of both long-range consequences of actions and anticipated future affordances. Goals at different levels of abstraction may, of course, conflict – for instance, resolving proximal interoceptive prediction error by eating chocolate cake might conflict with the longer-term goal of sticking to a diet (Pezzulo et al., 2018). Alternatively, temporary deviation from homeostatic set points at lower levels may be elicited to maintain higher level set points – such as a temporary change in blood pressure and adrenaline levels to engage fight-flight behaviour, with the goal of reaching safety and maintaining physiological integrity on a longer timescale. On the view of self-modelling in terms of allostatic control described, dimensions of the self at higher-levels constrain the self at lower-levels in that the self-model "actively shapes itself over time to align with those higher level regularities" (Hohwy & Michael, 2017, p. 370), for example long-term goals can be decomposed into intermediate short term-goals.

This section has explored how the self-model arises as a consequence of a sys- tem engaged in temporally deep active inference, as prior probabilities over particular policies depend on knowledge about what and where the system finds itself, and what actions are available to it (Friston et al., 2013; Moutoussis et al., 2014). Through active inference, agents can use their self-model to inform their goal and policy selection in order to arrive at high probability outcomes. This could entail assigning low probability to the selfoccupying states that are aversive, either physically or socially – with different hierarchical levels of the self-model contributing to different goal states. In this way, the hierarchical self-model determines salience—where "salience is literally defined by whatever has the most (or least) impact on visceral and autonomic homeostasis" (Allen & Tsakiris, 2018, p. 7), at increasingly deep spatiotemporal scales and levels of abstraction.

## 4. Psychedelics

One of the most striking and philosophically interesting effects of psychedelics is the radical disruptions of self-consciousness they can occasion (Huxley, 2010; Leary, Metzner, & Alpert, 1964), including apparently "selfless states" (Lebedev et al., 2015; Nour, Evans, Nutt, & Carhart-Harris, 2016). These states, instances of "Drug-Induced Ego-Dissolution" (DIED) are characterised by an experienced loss of self and/or loss of self/world boundary (Millière, 2017; Millière, Carhart-Harris, Roseman, Trautwein, & Berkovich-Ohana, 2018). DIED occurs most reliably under high doses of "classical" psychedelic drugs (5-HT2A receptor agonists), such as dimethyltryptamine (DMT), lysergic acid diethylamide (LSD), and psilocybin. Ego-dissolution appears to be induced more reliably under psychedelics than meditation, in a dose-dependent manner, and prompted most reliably by high-doses (Nour et al., 2016). Recent theoretical work has explored the phenomenological and neurophysiological similarities and differences of ego-dissolution induced by drugs and meditation (Millière et al., 2018; see also Limanowski & Friston, 2020, Millière, 2020, Sebastián, 2020).

### 4.1 Psychedelic therapy

Recent years have seen a resurgence of interest in the therapeutic potential of psychedelics. Several studies have found preliminary evidence that with administration in controlled circumstances psychedelics can be both safe and therapeutic, with an emphasis on the importance ofcontext in achieving therapeutic outcomes (Carhart-Harris et al., 2018). Interestingly, the positive therapeutic effects seem to scale with "peak" or mystical experience in the psychedelic state (Roseman, Nutt, & Carhart-Harris, 2018). Psychedelics have been shown to be effective in treating depression (Lyons & Carhart-Harris, 2018; Palhano-Fontes et al., 2019), obsessive-compulsive disorder (Moreno, Wiegand, Taitano, & Delgado, 2006), end of life existential distress (Griffiths et al., 2016), and have even been proposed as a potential treatment for disorders of consciousness such as the vegetative state and the minimally conscious state (Scott & Carhart-Harris, 2019). Carhart-Harris interprets the therapeutic effects as a result of "relaxing" high-level beliefs, allowing for a revision of pathological beliefs that have become overly dominant and resistant to revision, coined the "TIghtened BEliefs in Response to uncertainty" (TIBER) model (Carhart-Harris, 2019). The basic tenet is that under conditions of uncertainty the model falls back on "tightened" belief structures as a defence mechanism against intolerable stress and uncertainty. This fits with what we might expect under the FEP, as adopting shallow policies (such as addictive behaviours) may appear adaptive in the short term, rather than risking policies with greater expected free energy due to low precision or uncertainty. It has recently been proposed that psychedelics "relax" high-level priors in the generative model, allowing for the (context-dependent) revision of pathological high-level beliefs (Carhart-Harris & Friston, 2019). Both psychological insight and peak-experience in the psychedelic state appear to be predictors of long-term positive prognoses (Roseman et al., 2018).

### 4.2 Psychedelics in the predictive brain

The REBUS – "RElaxed Beliefs Under pSychedelics" – model of psychedelic function, offers a preliminary but promising model of psychedelic action where psychedelics, through 5-HT2A agonism, "relax" high-level priors or beliefs (Carhart-Harris & Friston, 2019). Here, the focus will be on how this mechanism may be cast under the hierarchical predictive processing framework as modulating precision-weighting. To bring this into focus, this section will review how precision-weighting sets a variable Bayesian learning rate in order to highlight certain features relevant to understanding the effects of psychedelics within this framework. Christoph Mathys and colleagues have recently developed a mathematical tool for modelling Bayesian inference modulated by expectations of volatility known as the hierarchical Gaussian filter (Mathys, Daunizeau, Friston, & Stephan, 2011; Mathys et al., 2014). The hierarchical Gaussian filter Mathys posits allows a system to optimally balance the influence of prediction errors in changing environments – in other words, to adjust its learning rate.

### 4.3 Bayesian learning rate

To recap, the predictive processing framework asserts that the brain instantiates "generative models" of the causes of incoming sensory data, iteratively updating these predictive models in light of incoming "prediction error" (Clark, 2013). This predictive inference is thought to occur across a hierarchy of inferred causes, where high levels track causes and regularities operating over deeper spatial and temporal scales, and lower levels track regularities over shallower spatial and temporal scales (Kiebel et al., 2008). The picture of the living or cognitive system as one which needs to optimise its own learning rate emerges out of the operationalization of Bayesian inference in predictive processing, namely in terms of predictions and precision-weighted prediction errors (Mathys et al., 2014). According to predictive processing the prediction is given by the prior probability (which itself comes from the previous posterior) and the prediction error is given by the difference between the prediction and the incoming sensory evidence. Prediction error is weighted according to the relative precisions of the prior and the prediction error (where precision is equivalent to the inverse variance of each probability distribution). Intuitively, highly precise prediction error will drag the posterior closer to the distribution of the sensory evidence and further from the prior, and in cases of low precision weighting of the prediction error, the inference relies more on the prior. This determines the learning rate:

> The more certain we are that the prior hypothesis is correct, the less we should be influenced by the prediction error (the evidence), which means that the learning rate is low. Conversely, the better the precision on the prediction error, the higher the learning rate; that is, the more we trust the quality of the evidence the more we should learn from it (Hohwy, 2017, p. 76)

In other words, the lower the learning rate, the greater the influence of top-down modulation from priors; the higher the learning rate, the greater the influence of the sensory evidence on the resulting posterior. Here, precision-weighting is the key mechanism – heavily weighted prediction errors drive a higher learning rate. In order to approximate Bayesian inference over time, it is essential for sensory systems to balance the learning rate appropriately. Over-reliance on priors means the system will fail to learn from experience, whereas over-reliance on sensory evidence (which may be noisy) will lead the system to "overfit". On this picture, Bayesian perceptual inference that minimises prediction error on the appropriate timescale – that is, not overfitting or underfitting the model – needs to have a means of regulating the learning rate (Mathys et al., 2014). This is implemented by building models of precision, or expected uncertainty, where higher-level priors track longer-term regularities that inform the relative precisions of more basic priors (Hohwy, 2017). Optimising the learning rate, and in-so-doing minimising prediction error over time, is a critical challenge the brain faces. This is equivalent to selecting a time frame over which to minimise prediction error. Minimising prediction error over too short a timescale—overfitting—runs the risk ofincreasing prediction error in the long run. Conversely, failing to accommodate new evidence will lead to underfitting, a failure to update predictions in light of new sensory evidence.

### 4.4 Psychedelic action as high Bayesian learning rate

In line with the REBUS model (Carhart-Harris & Friston, 2019), the relaxing of high-level priors under classical (serotonergic) psychedelics5 means the system adopts a very high Bayesian learning rate – that is, it is in a highly plastic state, in accordance with research showing an increase in plasticity under psychedelics (Ly et al., 2018). This picture casts the perceptual effects of psychedelics – "tripping" – as rampant overfitting of the sensory data, resulting from a loss of the usual constraint exerted by higher-levels on lower-levels of the inferential hierarchy. This "rampant overfitting", resulting from diminished influence from contextualising high-level priors tracking regularities on longer timescales, means the model fits a very short temporal scale, rapidly cycling through candidate models to account for the incoming sensory signal. It is worth highlighting a compatibility of the high Bayesian learning rate approach with other accounts of the mechanism of action of psychedelics in the predictive brain. The REBUS model posits the mechanism of action of psychedelics as reduced precision at high levels rather than increased precision at the sensory peripheries, as psychedelics appear to disrupt functioning via stimulation of 5-HT2A receptors on deep pyramidal neurons, thought to encode high level priors or beliefs (Beliveau et al., 2017; Carhart-Harris & Friston, 2019; Jakab & Goldman-Rakic, 1998). In contrast, Philip Corlett and colleagues have suggested that psychedelics preserve normal priors and act by increasing sensory noise through enhanced AMPA signalling (Corlett, Frith, & Fletcher, 2009; Corlett, Honey, & Fletcher, 2016). On this approach, if the relaxation of high-level priors is indeed an effect of psychedelics, it could be understood to be the result of the fact that "the persistence and strength of the sensory signal suggest that there is something to be explained" (Corlett et al., 2009, p. 521). Arbitrating between these two mechanistic accounts and disentangling causation – whether the relaxation of high-level priors causes the reduction in sensory gating, or reduction in sensory gating eventually lowers precision at high levels – becomes very difficult here, and it is not clear a simplistic causal account is the right approach. While identifying the mechanisms of action is a key empirical and theoretical project, one potential advantage of the high Bayesian learning rate hypothesis is that it doesn't distinguish between high precision at low levels and low precision at high levels, and as such remains agnostic over the mechanism of action while preserving the useful theoretical features of both accounts that will inform the theoretical account of ego-dissolution that follows.

### 4.5 Evidence for the high Bayesian learning rate hypothesis

A high Bayesian learning rate is concordant with the enhanced neural plasticity observed in individuals in a psychedelic state (Barre et al., 2016; Berthoux, Barre, Bockaert, Marin, & Bécamel, 2018; Ly et al., 2018). While an impairment to high-level cognition is found under psychedelics (Bayne & Carter, 2018), in line with the high Bayesian learning rate hypothesis, low-level learning (including extinction learning) and processing appears to be unaffected or enhanced in the psychedelic state (Carhart-Harris & Nutt, 2017; King, Martin, & Seymour, 1972; Romano et al., 2010). Further evidence for a high Bayesian learning rate under psychedelics is provided by a study looking at the effect of psilocybin on Kanisza triangles – perceptual objects where the brain "fills in" illusory contours using prior expectations – which found reduced filling in and a reduction in the related evoked potentials (Kometer, Cahn, Andel, Carter, & Vollenweider, 2011), concordant with the fact that a high Bayesian learning rate will reduce the effect of sensory history on current perception. In binocular rivalry studies – where different images are presented to each eye simultaneously, and are typically experienced as switching from one percept to the other – reduced switch rates and increased likelihood of the percept being a fusion of the two images has been observed under psilocybin (Carter et al., 2007, 2005), suggestive of less influence of priors on constraining current perception. Oddball paradigms are also suggestive of a weakened influence of priors on perception under psychedelics. In a sequence of tones, an "oddball" tone (unexpected given prior experience and context) generates a "mismatch negativity", an evoked brain response which has been interpreted in predictive coding terms as prediction error violating the expectations of the sequence (Garrido, Kilner, Stephan, & Friston, 2009). Under LSD, the surprise response to oddball stimuli is blunted, suggestive ofa weakened influence of prior expectations (Timmermann et al., 2018).

Arguably, there is also phenomenological evidence for the high Bayesian learn- ing rate hypothesis. Perhaps most eloquently articulated by Aldous Huxley: “Visual impressions are greatly intensified and the eye recovers some of the perceptual innocence of childhood, when the sensum was not immediately and automatically subordinated to the concept” (Huxley, 2010, p. 12). This observation lends itself to a straightforward translation into the terms of predictive processing, where “subordinated to the concept” can be understood as “constrained by higher-level priors”. More generally, psychedelic phenomenology such as dynamic distortions of spatial dimensions, where things change dramatically in size and shape can be understood as a failure ofhigh-level priors to canalise and constrain lower level predictions.

Psychedelic-induced ego-dissolution in active inference Given this picture of the action of psychedelics within a predictive processing framework, and the characterisation of self-models in terms of allostatic control, how should states of psychedelic-induced ego dissolution be conceptualised? The proposal here is that a loss of precision on high-level priors results in a flattening of temporal depth of the affordance landscape for the organism – precisely because it is high-level priors tracking longer timescales that structure temporally deep generative models. Recall, under active inference, lower and higher hierarchical levels encode regularities that unfold at faster and slower timescales respectively (Kiebel et al., 2008), such as the expected consequences ofaction both for proximal and distal goals (Pezzulo et al., 2015, 2018). Adopting a high Bayesian learning rate is equivalent to changing the time frame over which prediction error is minimised to fit very short timescales. As a result, the deep temporal models that typically guide action and policy selection collapse, and the faster timescales corresponding to lower levels are modelled in a much finer degree of detail (Pink-Hashkes, Rooij, & Kwisthout, 2017). On the account presented in this chapter, the self-model is constructed and bolstered in relation to affordances in the environment on several interlocked timescales, where high-levels contextualise and canalise the levels below and allow for motivational orientation to action opportunities pertaining to distal outcomes. Under psychedelics, the relaxation of high-level priors and the corresponding high Bayesian learning rate results in a collapse in the temporal thickness of deep generative models, and a collapse in the temporal depth of the corresponding self-model, which is understood as being is bolstered according to counterfactually rich expectations of the consequences of action on multiple timescales.

The collapse in temporal thickness can be understood as occurring due to a failure of sensory attenuation, occurring due to low precision at high-levels and a correspondingly high Bayesian learning rate. Similar stories about aberrant precision at high-levels of the hierarchy corresponding to inferences about affordances and agency have been proposed to underpin hallucinations and delusions in psychosis (Adams, Stephan, Brown, Frith, & Friston, 2013; Fletcher & Frith, 2009; Sterzer et al., 2018). Distinguishing between endogenous and exogenous causes—that is, distinguishing between perceptual inputs caused by oneself and those caused by the world—is vital for an agent to be able to effectively move through action space. Corollary discharges—predictions about the sensory consequences of actions—allow the system to do this by withdrawing precision from self-generated movements, and are thought to underpin experienced agency of actions (Crapse & Sommer, 2008; Friston, 2012b). The failure to predict the consequences of movement due to a failure of sensory attenuation is thought to result in an inability to attribute agency (Adams et al., 2013; Brown, Adams, Parees, Edwards, & Friston, 2013); for instance, a failure of corollary discharge has been thought to cause the attribution of inner speech to an external source in voice-hearing (Ford, Gray, Faustman, Roach, & Mathalon, 2007; Ford & Mathalon, 2005; Heinks-Maldonado et al., 2007). Importantly, for present purposes, corollary discharge can be understood as a kind of prior (Friston, 2010), and low-precision priors have been associated with the severity in psychotic symptoms and disturbances of agency in people with schizophrenia (Rösler et al., 2015). A reduction of precision on high-level priors in the psychedelic state means that the corollary discharges that would usually cancel out the expected consequences of actions fail to do so, generating an increase in prediction error at lower levels. These unexpected consequences are then attributed to external rather than internal causes, as the more prediction error is generated, the more likely an action (or thought) has exogenous rather than endogenous causes (Frith, 2003). This echoes similar themes in the autism literature. In autism, the failure of sensory attenuation "leads to the hypervigilant attention to sensory detail at the expense of a hierarchically deep explanation for sensations" (Picard & Friston, 2014, p. 1116) leading to what has been termed a "loss of central coherence" (Frith, 2003). Attribution to exogenous rather than endogenous causes could result in a loss of "perceptual mineness" – the background feeling that my experiences are "mine" – if, as has been argued, perceptual mineness is underpinned by anticipation of changes in perceptual inputs in relation to movements (Hohwy, 2007).

Ego-dissolution is not, however, confined to a loss of agentive control over immediate action outcomes, but may be characterised by a more profound dissolution of the sense of being a self or

"I" distinct from the outside world. On the view presented in this chapter, pre-reflective self-consciousness arises not just through modelling control over the most immediate sensory consequences of actions, but is bolstered by inferences about endogenous control over the distal sensory consequences of allostatic action and action policies. Under a high dose of a psychedelic, the temporary suspension on the gating mechanism on incoming sensory data, described in this chapter in terms of a high Bayesian learning rate, render both the proximal and distal sensory consequences of actions highly unpredictable, and the system ceases to have the sense of their being an agent which can (and should) be controlling sensory outcomes. Several authors have emphasised the psychedelic experience is a dynamic process as opposed to a firmly designated state (Masters & Houston, 1966; Preller & Vollenweider, 2016), and different types of ego-dissolution might occur both over the course of the experience and at different dosages. For example, inferences on the boundaries of the body (Blanke & Metzinger, 2009) might be increasingly blurred due to a failure to attenuate the flurry of low-level prediction error. Aspects of the self-model corresponding to longer timescales may break down due to a sustained failure of high levels to attenuate prediction error from low levels due to highly volatile prediction errors, consistent with the fact that bodily ego-dissolution tends to precede dissolution of narrative self (Savage, 1955). This fact is also perhaps suggestive, in opposition to the high-levels posited by the REBUS model, that ego-dissolution could be seen as the result of the high-levels failing to contextualise the upsurge of prediction error from across the cortex. The fact that the highest level of the self-model are "increasingly abstract, complex and invariant" (Limanowski & Friston, 2018, p. 5), may explain why higher levels of the self-model are going to be less perturbed by prediction error and perhaps only reliably altered at high dosages. Empirical exploration of these possibilities might be a fruitful avenue for future work, in particular through bridging the neurocomputational mechanisms posited here to both the dynamics of the experience as uncovered through "microphenomenological" interviews (Millière, 2017; Petitmengin, 2006), and to the underlying neural correlates of the experience (Timmermann et al., 2019).

The account of ego-dissolution in terms of a collapse in the temporal thickness of the affordance landscape presented here should also apply to the concept of a "cognitive affordance" landscape, where the "central function of autonomous activity in the mind wandering network is to create a constant stream of affordances for cognitive agency, a continuing internal competition among possible cognitive actions" (Metzinger, 2017, p. 2). Metzinger argues that mental actions – such as the volitional control of endogenous attention, or retrieval of an episodic memory – have epistemic rather that pragmatic goal states. On the allostatic control model of selfhood, the self-model would be constructed and bolstered relative not only to the expectations of the control of the sensory consequences of actions, but also the consequences of mental actions, where the consequences of a mental action might be epistemic and also interoceptive (consider a case where a memory triggers an autonomic response which subsequently acts as the afferent input to an interoceptive inference underpinning a felt emotion). Under psychedelics, loss of control of the expected outcomes of mental actions (as well as a loss of the pragmatic concerns usually driving which epistemic actions to take) might then also be fundamental to the experience of ego-dissolution. This idea is consistent with the fact that under psychedelics mental phenomena "take on the character ofobjective reality" (Savage, 1955, p. 12), where the ownership ofmental phenomena seems to subside and "the individual may feel like a bystander watching the mental activity of another person" (Girn & Christoff, 2018, p. 145).

It is worth mentioning a potential implication of this view for consciousness science more broadly. The psychedelic experience and ego-dissolution are often described as an "expansion" of consciousness. Friston (2018) argues that not only self-consciousness, but consciousness itself, is underpinned by temporal thickness: "consciousness is nothing more than inference about my future; namely, the self-evidencing consequences of what I could do" (Friston, 2018, p. 1). States of ego-dissolution, understood as collapse in the temporal thickness of the generative model, suggest that while temporal thickness very much structures our normal waking experience, it is not clear that temporal thickness ought to be equated with consciousness per se (see also Metzinger, 2020; Sebastián, 2020).

## 5. Ecstatic ego-dissolution and challenging experiences

The question remains as to why the hypothesised collapse in the temporal thickness of the self-model under psychedelics can be both ecstatic and of enduring therapeutic value. To bring this into focus, it's worth recapitulating core features of the self-model provided earlier. Recall, interoceptive inference on states of the embodied self "attunes" organisms to their affordance landscape, where inferences about the state of the embodied self (e.g. hunger) prescribe certain prediction error minimising policies (e.g. finding food). Inferences pertaining to allostatic consequences on longer timescales may mean higher-level imperatives trump lower-level drives, such as choosing to abstain from chocolate cake to stay healthy (Pezzulo et al., 2018). In the case of basic bodily needs, as described, these variables are controlled (Seth, 2015) through action – active inference is deployed to bring the world into line with predictions, rather than adjusting predictions (via perceptual inference) to conform to the world – for instance eating when hungry (Pezzulo et al., 2015). In just the same way that a hungry organism can act so as to harvest confirmatory evidence for the hypothesis "I am sated", hypotheses relating to higher-levels of the self-model geared towards control of outcomes on longer timescales act to constrain action in the present to bring downstream outcomes closer in line with the prior expectation. Overly precise priors driving action on a long timescale which are failing to be fulfilled, on this view, would be a persistent cause of suffering, due to the system consistently failing to meet (or align actions towards) the goal state (Hesp, Smith, Allen, Friston, & Ramstead, 2019). Under the model of psychedelic-induced ego-dissolution proposed, the high-precision high-level priors geared towards control on multiple timescales cease to exert influence on the system due to the proposed lowering of precision of high-level priors under psychedelics. If action ordinarily arises from a process of minimising deviations between the organism's actual (inferred) and desired trajectory (Friston et al., 2010), the loss of precision on high-level priors means that, instead of driving action policies, they lose influence on the rest of the system and cease to structure pre-reflective self-consciousness to orient to action opportunities favouring their fulfilment. As these prior beliefs are relaxed, they instead become amenable to perceptual revision from the influx of (highly precise) interoceptive and exteroceptive information. The collapse in temporal depth in the psychedelic state is therefore not experienced as a loss of allostatic control, precisely for the reason that the priors pertaining to longer timescales are no longer asserting an influence on the system and constraining action (and perception) in their usual manner. This picture seems to align well with phenomenological reports of ego-dissolution: "It felt as if 'I' did no longer exist. There was purely my sensory perception ofmy environment, but sensory input was not translated into needs, feelings, or acting by 'me' " (unpublished online survey data quoted in Millière et al., 2018, p. 7). Peak experiences under psychedelics, then, could be understood as absence of prediction errors relating to allostasis due to a flattening of the temporal depth of the affordance landscape, resulting in the feeling of "oceanic boundlessness" – a sense of immense well-being and peace. Here, the "itinerant strategies" to stay within our "species-specific window of viability" (Clark, 2013, p. 13), are no longer necessary as the "first prior" – the expectation or imperative for existence – is being met without conditions.

Following the TIBER model, many psychopathologies may be due to high precision on high-level priors (Carhart-Harris, 2019; Clark, Watson, & Friston, 2018). Peak psychedelic experience may act as a "reset" allowing for revision of entrenched high-level beliefs that structure pre-reflective self-consciousness (and, accordingly, the affordance landscape) – opening up new domains of salience and possibility for meaningful engagement with the world, through revised and retuned self-models. Increased bottom-up information flow (particularly from the limbic system), through a high Bayesian learning rate, may make entrenched high-level priors amenable to revision via perceptual inference rather than driving control via active inference. This lays the theoretical groundwork for why psychedelics may effectively treat depression: if depression is underpinned by a high precision prior of low allostatic self-efficacy (Stephan et al., 2016), it follows that relaxation and revision of this prior should alleviate depressive symptoms. Finally, (and speculatively), if the account of "retuning" of self-models under psychedelics presented here generalises to the bodily self (which the experiential changes in bodily selfhood would suggest) this account is suggestive of a potential role for psychedelics in the treatment of chronic pain, and for phantom limb pain—for which there has already been promising results (Fanciullacci, Bene, Franchi, & Sicuteri, 1977; Ramachandran, Chunharas, Marcus, Furnish, & Lin, 2018). The primary focus so far has been on "peak" experiences, due to the growing number of papers indicating they are central to positive long-term therapeutic outcomes (Roseman et al., 2018). However, while generally psychedelics are thought to be very low risk (Nutt, King, & Phillips, 2010), and there is evidence to suggest they are protective against mental health problems (Hendricks, Thorne, Clark, Coombs, &Johnson, 2015), acute and occasionally persistent adverse psychological reactions do sometimes occur (Strassman, 1984). While "complete" ego-dissolution is described as a "state of complete surrender, associated bliss, and union with all things" (Carhart-Harris & Friston, 2019, p. 321), "incomplete" ego-dissolution – due to psychological resistance or an insufficient dose – can be characterised by intense fear, anxiety, or distress. On the account presented in this chapter, this can be understood as resulting from psychological resistance, where psychological resistance here may be conceptualised as a high-precision prior on being able to control the experience, that is maintained though fear-driven endogenous attention. Failure to control the experience, in violating the highly precise prior for the goal state of control, is then experienced as a loss of allostatic control, bringing with it feelings of intense fear or distress. In therapeutic contexts, encouraging users to "let go" and "surrender" to the experience (Richards, 2015), could be understood in these terms, as discouraging the user from putting high (endogenous) precision on a prior for control that could result in adverse experiences when unfulfilled. These considerations highlight the essential importance of context in achieving therapeutic outcomes (Carhart-Harris et al., 2018).

## 6. Conclusion

Psychedelics are known for their ability to profoundly alter consciousness and occasion so-called "mystical" experiences (Huxley, 2010). The renaissance in psychedelic research in the past decade is beginning to shed light on the mechanisms underpinning the extraordinary states of consciousness induced by psychedelics (Carhart-Harris & Goodwin, 2017). Within psychedelic phenomenology, experiences of ego-dissolution are of particular phenomenological, philosophical and therapeutic interest (Letheby & Gerrans, 2017; Millière, 2017; Nour & Carhart-Harris, 2017). This chapter has given a preliminary account of how ego-dissolution under psychedelics can be understood in terms of predictive processing and active inference. The hypothesis here is that the action of psychedelics within the predictive processing framework is best understood as a "relaxation of high-level beliefs" (Carhart-Harris, 2019), and this can be unpacked in terms of a high Bayesian learning rate (Hohwy, 2017; Mathys et al., 2014). Psychedelic-induced ego-dissolution, then, results in a collapse in temporal thickness (Friston, 2018) of the self-model as conceived within an active inference framework. The therapeutic effects of ego-dissolution, then, can be understood in terms of the relaxing and retuning of entrenched self-models, or a "resetting" or "opening" of the affordance landscape, allowing for the possibility of new modes of engagement with the world, oneself, and other people.

# Chapter 4: Losing Ourselves: Active Inference, depersonalization, and meditation

Disruptions in the ordinary sense of selfhood underpin both pathological and 'enlightened' states of consciousness. People suffering from depersonalization can experience the loss of a sense of self as devastating, often accompanied by intense feelings of alienation, fear and hopelessness. However, for meditative contemplatives from various traditions, "selfless" experiences are highly sought after, being associated with enduring peace and joy. Little is understood about how these contrasting dysphoric and euphoric experiences should be conceptualised. In this chapter, we propose a unified account of these selfless experiences within the active inference framework. Building on our recent active inference research we propose an account of the experiences of selfhood as emerging from a temporally deep generative model. We go onto develop a view of the self as playing a central role in structuring ordinary experience by 'tuning' agents to the counterfactually rich possibilities for action. Finally, we explore how depersonalization may result from an inferred loss of allostatic control, and contrast this phenomenology with selfless experiences reported by meditation practitioners. We will show how, by beginning with a conception of self-modeling within an active inference framework, we have available to us a new way of conceptualizing the striking experiential similarities and important differences between these selfless experiences within a unifying theoretical framework. We will explore the implications for understanding and treating dissociative disorders, as well as elucidate both the therapeutic potential, and possible dangers, of meditation.

## 1. Introduction

In daily life we take for granted the existence of a self: we feel we are possessors of certain qualities, the experiencers of certain sensations, that we are different and distinct from one another and that we endure from day to day. And yet these assumptions have long been the focus of skepticism within both Western and Eastern philosophical traditions. Thinkers from various disciplines (e.g. from philosophy of mind, cognitive science, phenomenology and Buddhist philosophy) are beginning to collaborate on various topics revolving around self and subjectivity. One lens through which philosophers and cognitive scientists have been recently exploring the self is through cases where subjects report a loss, or diminishment, of their sense of self. These reports occur most prominently in the context of psychiatric disorders such as depersonalization (e.g. Miller et al. 2020; Seth et al. 2012; Colombetti & Ratcliffe, 2012), meditation (e.g. Lutz et al. 2019; Britton, 2019) and psychedelic drugs (e.g. Deane, 2020; Millière, 2017). A particularly promising framework for approaching and understanding these phenomena is the active inference framework, a popular approach to action and perception that uses principles of variational Bayesian inference (Friston, 2017).

Our aim in this chapter is to provide an updated account of selfless experience within the active inference framework.[4] By selfless experience, here, we mean the diminished sense of self that is reported in a wide variety of cases including depersonalization and meditative insight. Active inference and predictive processing have already been used to provide accounts of depersonalization in psychiatric contexts (Seth, Suzuki and Critchley 2011; Gerrans 2019), and although we find these accounts promising, we seek to build on them in important ways. In particular, we differ from existing accounts in taking affective valence and control to be central to the sense of self. Building on existing accounts of self-modelling within an active inference framework (Friston, 2018; Hohwy & Michael, 2017; Seth, 2014), our account casts the self-model in terms of an allostatic control model (ACM; Deane, 2020), which we unpack in terms of 'agentive control' and 'motivational' components. The central thesis of this view is that the self is understood as an inference about endogenous causes of self-evidencing outcomes. In simple terms, this could be understood as the system modelling what it *wants* (motivations) and what it *can do* (abilities). However, we do not simply adopt the ACM for its own sake — there are concrete explanatory pay-offs. In particular, we are better able to account for the wide range of selfless experiences under a single unifying framework. Selfless experiences come in a variety of flavours, ranging from the dysphoric and dysfunctional experiences associated with depersonalization, to the euphoric and potentially super-functional states sought after by meditators. Our explanation of this difference is, as we will see, an intrinsic part of our account of the emergence of these phenomena themselves.

An important addition to this literature that we will make is a reinterpretation of the role that affect plays in these processes. We will argue that the sense of self arises from the system's evaluation of its

[4] See Ciaunica et al 2020 for a recent phenomenological account of depersonalization and meditative insight.

own performance, or predictive control, of its own adaptive behaviours. As we will see the tracking of our performance, and the allocation of resources (i.e. setting of precision), is being done in part by affective systems. That is, we quite literally feel how well adapted we are to a situation, and those feelings move us in ways that are intended to improve that fit. This has the consequence that our sense of being a self and affect are mechanistically intertwined. This updated theoretical account of selfhood then allows us to propose a more unified framework for understanding various alterations in selfhood and affectivity.

We proceed as follows: In section 2 we give an overview of the free energy principle and hierarchical predictive processing. In section 3, we position these frameworks within a control theoretic perspective, and show how allostasis can be formalised in terms of active inference. In section 4, we build on these ideas to present an account of self-modelling in terms of allostatic control. In sections 5 and 6 we apply this model of the self to address depersonalization and selfless experiences attained through meditation respectively. We wrap up and conclude by comparing and contrasting these two dysphoric and euphoric selfless experiences.

## 2. From the Free Energy Principle to Hierarchical Predictive Processing

The free energy principle (FEP; Friston, 2010) is an ambitious unifying and overarching theory of life, according to which biological systems naturally strive to minimize free energy.

The FEP starts from the observation of existence (Friston and Stephan, 2007; Friston et al., 2010) and seeks to understand how organisms maintain their existence by 'tuning' to their environmental niche, where the quantity of free energy is understood as a measure of the disattunement (which is equivalent to model "uncertainty") between the agent and environment (Bruineberg and Rietveld, 2014). Crucially, in order to exist and reproduce, agents must stay within conditions that are conducive to continued existence – such as avoiding an unacceptably high body temperature. Of course, this is *phenotype-specific* – the conditions that make continued existence viable vary across species. Organisms must minimise free energy, which is equivalent to maximizing the evidence of their model, and so their own existence (Hohwy, 2016; Friston 2010, 2013). Maximising model evidence in this way is called 'self-evidencing' (Hohwy, 2016).

In animals like us (and many others) it has been proposed that free energy is minimized, at least in large part, by hierarchical predictive processing in the brain and central nervous system (Friston, 2005; Clark, 2013). What the brain has to do, on such a view, is minimize prediction error (free energy) as efficiently as possible. This requires it to come up with an overall hypothesis or model about what is going on in the world. This hierarchical model generates predictions and, if it is inaccurate, it generates *prediction error* and updates predictions accordingly.

A major challenge in model selection arises because the world is a noisy and ambiguous place. So, there exists, at any given time, more than one model that fits the incoming sensory signal. This is where the notion of *prior probability*, often shortened simply to *prior*, comes in (and with it the Bayesian element of the framework). This is the background probability of the model independently of the evidence. For example, (adapting an example from Pezzulo, 2014) before I hear my downstairs front window creak open, there is a background probability concerning the likelihood that I might be burgled. Whether I live in a high or low crime neighbourhood will influence the prior probability of the "that's a burglar!" model in response to the sound of the creaking window ("the evidence"). Models are selected based on both fit with current evidence and their prior probability. This means that you can get trade-offs, for example, where a model with a relatively low fit has a sufficiently high prior probability to be selected. For example, in the case of the hollow mask illusion, the model with the best fit would be the (perceptually accurate) "hollow concave face" model, but the slightly lower fit "normal convex face" model has such a high prior probability that it is selected instead, giving rise to the illusion.

This captures what the brain has to do, namely, resolve ambiguity using priors (viz. in a Bayesian manner), however, it doesn't tell us how this is implemented physically in the brain. Put simply, the brain maximises efficiency (minimizes free energy) by being proactive and anticipatory. In other words, the nervous system doesn't passively wait for inputs to come in. Rather, even at the earliest stages of sensory processing, inputs are greeted by a barrage of top-down prediction. This doesn't just save time, it also saves energy and bandwidth, since the parts of the incoming sensory signals that have already been accurately predicted don't need to be passed up the processing hierarchy. All that gets passed up is what is "newsworthy" (Hosoya et al. 2005), namely, prediction error. Putting this all together, the nervous system tries to minimize prediction error by coming up with successful hierarchical predictive models that are chosen in a Bayesian manner (namely based on fit and prior).

There are two more important tweaks to this picture. The first is to do with second-order prediction dynamics, namely, how the brain deals with statistical volatility. This requires introducing the notion of *precision*. In short, the world that we live in doesn't just have variability, but also predictable levels of variability. As a result, our nervous systems learn over time that there are contexts where environmental information is high quality (trustworthy), and other contexts where it's not. For example, in good lighting visual information is relatively high quality, whereas in poor lighting it's relatively low. What an optimal system will do in response to this is have a way of setting second-order precision, namely, of appropriately varying the extent to which prediction error should be taken seriously (adjusted as a function of the likelihood of prediction error being accurate or simply noise). In high-quality informational contexts, it is expected that predictions will be good, and so prediction errors will be given relatively high weight (or gain). In low-quality contexts, prediction errors will be taken less seriously. This turning up and down of the gain on prediction error signaling is most commonly called *precision-weighting*, and it plays a role far beyond the second-order dynamics that we used to introduce it. It is central to attention (Hohwy, 2012), and to the bringing about of bodily movement, an issue to which we now turn.

The second tweak comes when we note that, for embodied creatures like ourselves, action is an ever-present part of our existence. The Bayesian picture just described makes it look like we are primarily in the business of updating our models to best fit inputs from the world. But, of course, there are two ways of responding to prediction error. You can, certainly, update the model to better fit the world, but you can also update the world to better fit the model. The former is known as *perceptual inference* and the latter is known as *active inference*. It is with the latter that you get a PP account of action, and basic motivation more generally. Active inference, on our view, is central to allostasis, a notion we introduce shortly.

## 3. A Control-Theoretic Perspective

The foundations of active inference can be traced to control theory. The idea that a system maintains existence by resisting environmental disorder by acting to remain within a limited repertoire of *phenotype-congruent* states is closely related to the notion of maintaining 'essential variables' (Ashby, 2013), where an internal reference point (also known as a setpoint or goal signal) is compared to the current state, and the system acts so as to restore conditions to the setpoint.

The principles of *control-oriented predictive regulation* (Seth & Tsakiris, 2018) are very similar.
Here, the brain applies the same inferential machinery of hierarchical predictive processing to infer and track key homeostatic variables, using prior expectations and afferent sensory information about the body coming 'from within' (Craig, 2003).In order to stay alive, organisms have to execute the right actions to bring about state transitions that bring bodily states into reasonable bounds (Pezzulo et al. 2015). The phylogenetically endowed high-precision on expectations for staying within homeostatically viable states means that the organism acts to realise prior beliefs corresponding to the maintenance of essential variables ('goal priors'). For example, eating to restore a blood sugar concentration to expected levels. While goal priors originate in the maintenance of essential variables (e.g. steady temperature, blood sugar levels, etc.), over the course of ontogeny an organism can acquire new goal priors which are predictive on longer timescales of being relevant for maintaining homeostasis—such as staying within a particular social milieu (Matthew & Tye, 2019).

Active inference, then, formalises homeostasis through a control theoretic lens. Homeostasis from this perspective is maintained not only through autonomic reflexes (i.e. sweating to cool down), but also by *prospective control*. Such systems anticipate future dyshomeostatic conditions before they arise, and proactively act to avoid them. This prospective control relates to both inferences about current bodily states, and future bodily states contingent on certain actions (Pezzulo et al. 2015; Seth, 2014; Sterling, 2012). This process of anticipatory action, by which the brain regulates the needs of the body, is known as *allostasis* (Corcoran et al. 2020). Active inference formally articulates allostasis, such that agents *anticipate* surprising outcomes before they arise, and act in order to minimise uncertainty about potential future outcomes (Pezzulo et al. 2015, 2018; Sterling, 2012).

On the active inference formulation, the action selection process itself is cast as a problem of inference, where agents must infer the active sampling of the world which realises prior preferences and minimises uncertainty (Kaplan & Friston, 2018). Action selection, then, depends on the use of a deep temporal model, where policies (sequences of actions) are selected based on prior expectations of the quantity of free energy the agent expects itself to average over time ('expected free energy') *given* a particular policy or course of action (Friston et al. 2017; Pezzulo et al. 2015). Intuitively, some courses of action (such as riding in the train carriage) have lower expected free energy than others (such as riding on the roof). Crucially, this involves anticipating unfavourable or dyshomeostatic conditions before they arise, and acting to minimise uncertainty about potential future outcomes (Friston et al. 2015, 2016, 2017). On this account, higher levels of the cortical hierarchy, tracking regularities unfolding on longer timescales (Keibel & Friston, 2008), contextualise lower levels by anticipating the downstream consequences of action, and selecting policies which minimise expected free energy according to these expectations (Friston 2010; Pezzulo et al. 2015). For example, my longer term goal of successfully catching the train includes expectations about what I need to do to get to the station on time, which in turn unpacks into sub-goals such as getting in my car, and lower-level action-prediction loops as I use the pedal, gearstick, and so on.

In order to minimise free energy over longer timescales, active inference requires balancing the *pragmatic* and *epistemic* value of different actions. The pragmatic (or instrumental) value of an action or action policy (a sequence of actions) refers to the probability of it resulting in sensory states that fulfil some prior preference or goal state, such as maintaining a viable body temperature. Epistemic value refers to the reduction of uncertainty or information gain expected under a given action or action policy (Kaplan & Friston 2018). Epistemic action allows organisms to increase an agent's ability to reduce free energy by increasing their understanding of the predictable aspects of the environment. Information-seeking behaviour such as novelty-seeking and curiosity can be accounted for within this formulation in terms of epistemic action (Kiverstein et al. 2017; Friston et al. 2015; Kaplan & Friston 2018; Mirza et al. 2016, Pezzulo & Nolfi, 2019). Intrinsic motivation (and epistemic foraging) can be understood here in terms of uncertainty reduction (Barto et al. 2013). Simulations of economic decision-making and epistemic foraging behaviour have been built based on this view that the probability of a policy is proportional to expected free energy (Friston et al. 2014, 2015, 2017). Active inference formulations of planning and navigation have been used to dissolve the 'explore-exploit' dilemma, as the agent simply needs to act so as to minimise uncertainty (i.e. free energy; Kaplan & Friston, 2018). Agents engaging active inference don't just keep themselves in the states that are expected, rather they anticipate in order to minimise uncertainty about potential future outcomes (Friston et al. 2014, 2015, 2017; Schwartenbeck et al. 2013).

Now that we have introduced the control-theoretic notion of allostasis, and how it is achieved via active inference, we next go on to develop our view of the sense of self.

## 4. The Sense of Self as a Model of Allostatic Control

This section will argue that the self is best understood in terms of an *allostatic control model* (ACM). Recently, a number of computational models of the minimal sense of  self (namely, the self as implicitly present in everyday world-directed experience, rather than something more overt and explicit like the self-conception or narrative self) have been advanced in the active inference literature (Seth 2013; Limanowski & Blankenburg 2013; Apps & Tsakiris, 2014; Allen & Friston, 2016). Common to these proposals is that the sense of self arises inferentially within a hierarchical generative model. Our central claim is that the inferential self-model arises from the system tracking its own self-evidencing capabilities (Friston 2018). The purpose of tracking these capacities is to infer confidence (precision) in potential action policies according to their expected free energy, and thereby arbitrate between potential actions accordingly. Self-modelling of this kind, then, is fundamentally related to selecting allostatic or anticipatory actions, where the system preemptively infers and avoids unfavourable conditions before they arise. By casting the self-model in terms of allostatic control we will connect our view in new ways to the prevalent theme in neuroscience about the rich relationship between affectivity and the self (Allen & Tsakiris 2018; Seth 2013; Damasio 2003). This view can be understood formally in terms of an higher-level inference about 'subjective fitness' - that is, a higher level of the generative model that scores the 'fit' between the action model and the world (see Hesp 2020 for a formal treatment and computational model of this idea). Conceptually, our view of the sense of self can be decomposed into 'agentive control' and 'motivational' components. We will present these in turn.

### 4.1. Agentive Control

Recall that while perception involves updating the model to better predict the incoming sensory input, action changes the incoming sensory input to better fit the model. In selecting an action, then, the system implicitly infers itself as able to bring about the consequences of that action. A sense of agency, the sense of being the one in control of an action, naturally emerges here as part of model sampling (Friston et al. 2013) - in selecting an action the system implicitly infers itself as able to bring about the sensory consequences of the action. On this view, the sense of control – the expectation of being able to bring about certain consequences given certain actions – is *learned* through past experiences of the system inferring its own agentive capacities.

This connects closely with pre-existing accounts of the sense of agency, where the system infers its own agency based on the ability to predict the outcome of a given action (Haggard 2017). Here, attribution to endogenous causes (self), as opposed to exogenous causes (world/other), occurs as the result of a "comparator model", where the sensory consequences of an action are compared with the expected sensory consequences (Frith 2014). This allows the system to sculpt and improve motor control, as the discrepancy between the sensory consequences of an action are compared with the predicted (intended) outcomes. The system can then act to iteratively reduce this discrepancy and refine motor commands (Miall & Wolpert 1996; Wolpert & Flanagan 2001).

Crucially for the current account, *control* is *temporally deep* (Pezzulo 2018), such that the agent not only has predictions about the immediate consequences of actions, but also of consequences extending into the future. The sensory consequences of a given action may be sensorially proximal (e.g. the immediate sensory consequences of hitting send on an email), or sensorially distal and abstract (e.g. the expectation that when you see that person they will know the information in the email). The system, then, must be able to track the outcomes of actions on multiple timescales. Within the generative model, lower and higher levels of the hierarchy track regularities unfolding at faster and slower timescales respectively (Kiebel et al., 2008). For an organism with a temporally deep generative model this includes tracking its expected control of actions on short timescales (e.g. the expected sensory consequences of taking a step), and using these inferences to inform inferences about the state of control on temporally deep timescales (e.g. being able to walk a distance). On this view, the system models itself as an agent according to this hierarchically deep inference about its own endogenous control of sensation via its actions. In other words, I have a sense of what I can do based on past experience of acting in the world, and come to expect myself as a controller over my future actions.

### 4.2. Motivation

The *motivational* component of our view of the self-model is understood in terms of *goal priors,* and as such connects closely to views of selfhood grounded in interoception (Barrett 2017; Friston, & Seth 2016; Seth & Tsakiris 2018). This is because the system will be generally more concerned about controlling the 'essential variables' (i.e. homeostatic set points like blood sugar levels) tracked by interoception, than variables inferred through exteroception and proprioception, which are less likely to pertain directly to homeostasis (Seth 2014; Seth & Tsakiris, 2018).

Creatures tracking longer timescales can augment this with *deep goal hierarchies* (see Pezzulo et al. 2018), where fulfilment of longer-term goals can be traded off with fulfilment of shorter-term goals. On this view, low-level maintenance of 'essential variables' are phylogenetically endowed expectations that, due to an "a priori hyper-precision of visceral channels" (Allen 2018: 7), the system must act to fulfill, rather than simply updating via perceptual inference. One example would be moving to the shade under a tree to maintain viable body temperature. Divergence from these fundamental, phenotype-congruent low-level prior expectations tunes attention and amplification of sensory signals. This manifests itself to the system as, for example, the feeling of hunger (interoceptive prediction error), or a violation of the "healthy body condition" prior in the case of pain (Ongaro & Kaptchuk 2019). These interoceptive changes tune the organism to the appropriate action opportunities in the given context, such as finding food to resolve interoceptive prediction errors or removing the source of pain. Crucially, pain is tuned relative to expectations given the context (Moutoussis et al, 2014). In the case of an approaching bear, the prospective inference about imminent catastrophic prediction error of being eaten trumps the proximal pain of a twisted ankle, and the selected policy is running away. Put another way, hierarchically deep contextualization of interoceptive signals tunes an organism to appropriate actions and engagements with the environment (Pezzulo & Cisek 2016), and assigns appropriate precision to priors and ascending prediction errors. For low-level drives and motivations, this is intuitive – the hungry organism is tuned to capitalize on eating opportunities present in the environment. Precision on goals tracking different timescales are continually being traded off between levels - such as refraining from eating chocolate cake in the present for the sake of a longer-term goal of sticking to a diet (Pezzulo et al, 2018).

The sense of self, then, emerges as the result of a hierarchically deep inference about the system's control of its own self-evidencing outcomes. In the generative model, this means the sense of self can be understood in terms of a higher-level inference about the 'fit' between the current action-model and the world. By fit here we mean how well or poorly one is doing at reducing error over time relative to expectations. As we will see in the next section, a key implication of this picture is that self-modelling is fundamentally affective, where affective changes in the body tracks how well the organism is doing at fulfilling its own goal priors ('subjective fitness'; Kiverstein et al. 2017; Joffily & Coricelli 2013; Seth & Friston 2016). An upshot of this picture is that self-modelling, and indeed the feeling of being a self, is connected to affect in ways previously underappreciated in the literature, as we will explore next.

### 4.3. Affect in Deep Self-Models

The previous section argued for a view of self-modelling as a higher-level inference about the system's allostatic control. In our view, as we will now see, minimal self-modelling and affect are co-constitutive, such that affect can be understood as an inference about the performance of the action model in bringing about self-evidencing outcomes. This section unpacks how inference about allostatic control, manifesting affectively, is central to the allocation of precision.

In tracking the performance or 'fitness' of the model over time the system becomes sensitive to *the rate of error reduction.* In selecting a policy, the system has prior expectations of the rate at which error is likely to be reduced over time. The system can then evaluate whether its performance at reducing error is better or worse relative to its prior expectations. We can think of each agent's performance in reducing error then in terms of a slope that plots the various speeds that prediction errors are being accommodated relative to their expectations. Changes in the rate at which error is reduced (referred to as "error dynamics") turns out to be an important source of information for a predictive organism as it reflects the efficiency, and so the quality, of its action model performance over time. As such, error dynamics play an important role in tuning precision estimations – increasing or decreasing our beliefs in the reliability of the model generating the policy (Kiverstein et al. 2017; Hesp et al. 2019). If precision is set based on estimations of how likely some action is to lead to the expected result, then the efficiency - the rate at which error is reduced - of those actions to reduce error should be taken into consideration. Greater than expected error for a given policy is evidence that the system should down-regulate precision on the action model. Sensitivity to error dynamics increases our capacity to reduce prediction error over longer timescales, as it affords a means to toggle confidence levels on the action model according to the volatility of the environment.

The phenomenological manifestation of this (subpersonal) sensitivity to error reduction rates over time is affect. There is a growing literature that supports the view that affective changes not only track changes in immediate divergences from the homeostatic ideal, as was the focus of earlier predictive accounts of interoception (see Seth 2013), but also tracks the rate of change in error management over time (Kiverstein et al. 2017; Joffily & Coricelli 2013; Van de Cruys 2017). Valenced bodily feelings (i.e. positive and negative hedonic tone) are, in part, a reflection of how well or poorly we are reducing error over time relative to expectations. When error is being reduced slower than expected, and the organism is becoming increasingly disattuned to its environment, this change is marked by feelings of frustration and disappointment. The negatively valenced bodily feelings provide the organism with feedback about the reliability of the selected action policies, indicating a need to down-regulate precision on those policies. In contrast, when error is being managed at a better than expected rate the organism is gripping the scene well, the bodily feedback are positive feelings of hope and satisfaction, precision is up-regulated. It is intuitive that persistently worse than expected rates of error reduction on a given goal prior act as a disincentive to pursue that goal and motivate the system to select a more achievable goal, and doing well is motivating to continue to realise a certain goal. Precision doesn't just concern the organism here and now and its momentary state of uncertainty but is instead helping it to continuously improve working towards managing uncertainty over time. Importantly, positive and negative feelings alter precision relative to the rate at which we have come to expect errors to be resolved.

Affective valence here is being reimagined within the active inference framework as a domain general controller that tracks and assigns precision relative to changes in our expected rates of error reduction (that is, expected reductions in free energy; Kiverstein et al. 2017; Hesp et al. 2019). Inference about how well the system is self-evidencing as a whole is tracking a long-term dimension of the self which is necessarily more invariant and abstract in virtue of tracking a longer timescale, showing less variability than 'lower' aspects of the self-model that are more amenable to changing across contexts. Negative and positive feelings then track lesser than expected and greater than expected allostatic control respectively. This higher-order inference about the system's confidence in its own action model, used to modulate precision on expected free energy, (Hesp et al, 2020), is a candidate computational correlate for the sense of self – the feeling of being an agent.

This account of self-modelling as mechanistically intertwined with affectivity is, at present, a theoretical proposal. However, recent work (most notably Hesp et al 2020) provides proof of principle of how this theoretical framework can be modelled computationally. This is very promising groundwork for future work in computational modelling that is able to tie both phenomenology and behaviour to underlying computational mechanisms. An important consequence of highlighting this underappreciated link between affect and self-modelling in active inference is that it provides a bridge between these computational frameworks and the phenomenology of being a self (for another account of this see Kiverstein et al. 2020). Bodily feelings here represent a pre-reflective source of information about how well an agent is doing in their predictive engagements. These feelings give them a sense of what they can do, of what is possible and what is not possible (the sense of "I can"). We have a feel for what is possible in the world based on what we can do in the particular situation we find ourselves within. Above we characterized bodily feelings as driving policy selection. The result is that one quite literally feels drawn to relevant action possibilities. These bodily feelings track which possibilities are relevant to an agent, and move us to improve.[5] The result is an ongoing dynamic dialectic between agent and environment all circling around affectivity.

While the importance of this ongoing tension between bodily feeling and environmental affordances is easily overlooked when it is functioning well, alterations in this quality can have devastating effects on how one experiences oneself and one's world. With the addition of these more recent computational models of valence as setting precision relative to changes in control, we have now for the first time at our disposal the means to provide the fullest expression of an active inference account of the sense of self. In the rest of this chapter we will use this more fully realized view of the self to propose a new unified account of the alterations in self-experience native to depersonalization and meditative insight.

[5] For an excellent account of the neuroscience supporting the role of affect (including valence and arousal) in simultaneously tracking the relationship between the organism and the environment, and preparing the organism to make improvements to that relationship, see Lisa Feldman-Barrett's work (2017).

## 5. Active Inference Accounts of Depersonalization

Depersonalization disorder (DPD) is still a relatively neglected dissociative disorder. Dissociative disorders are a class of mental illness characterized by disruptions in perception, consciousness and/or identity. These disruptions can cause various symptoms that are problematic for a person's life including social relationships and work life. Recently, however, research into the phenomenon of depersonalization more generally is increasing in part due to the piqued interest of philosophers and cognitive scientists interested in the nature and function of the self. Depersonalization experiences potentially provide researchers with important glimpses into the neuropsychological mechanisms and functional profiles of our ordinary experiences of being a self (see Metzinger on philosophy and dissociative disorders). A hallmark of depersonalization is a disturbance in subjective experience. This commonly includes a sense of detachment or alienation towards themselves, their bodies and their environments. While specific disturbances in self-related experiences (depersonalization) and their experience of the environment (derealization) can come apart, they commonly co-occur, we'll have more to say about his co-occurrence shortly (Sierra and David, 2011).

London-based writer Gracie Lofthouse writes on her own experience of depersonalization in a recent article:

> "The first time I can remember feeling like I didn't exist, I was 15. I was sitting on a train and all of a sudden I felt like I'd been dropped into someone else's body. My memories, experiences, and feelings—the things that make up my intrinsic sense of "me-ness" — projected across my mind like phantasmagoria, but I felt like they belonged to someone else. Like I was experiencing life in the third person" (Lofthouse 2014).

Most people have some experience of this sort of state. If you haven't, it can be difficult to understand, and indeed sufferers of depersonalization commonly report difficulties in expressing their experiences (Simeon & Abugel, 2006, p. 80). Depersonalization symptoms can last for moments, or several years, and commonly accompany major depression, anxiety disorders, substance addiction, brain injury and disease, and emotional trauma. An increasingly popular view of depersonalization is that it may act like an "air-bag" in traumatic situations: when fight or flight are unable to remove an overwhelming, emotionally-painful, experience then the affective system may have its volume turned down as a direct means of reducing the suffering. The result of this reduction is what Medford calls it "desomaticion" or "deaffectation" (Medford, 2012; Medford et al. 2016; Sierra et al., 2005; Simeon et al., 2008), and is potentially the cause of the characteristically strange phenomenon of losing something important about the self and the world (see, e.g., Baker, Hunter, Lawrence, & David, 2007; Medford, Sierra, Baker, & David, 2005; Radovic & Radovic, 2002; Sierra, Baker, Medford, & David, 2005; Sierra & Berrios, 1998; Simeon & Abugel, 2006).

In our view, predictive processing has offered some of the most promising avenues for understanding depersonalization. Our main aim here is to build on these, and to improve on them based on more recent developments in the literature on active inference and the view of the self-model outlined in the previous section. The main explanatory payoffs that we can see are, not only that we can better explain depersonalization and related symptoms, but that we are also well-placed to explain why some instances of loss of self can have a positive valence, while others do not. Ultimately, superficial similarities in what are described as experiences of "loss of self" mask deep underlying differences.

### 5.1 Existing Predictive Processing Accounts of Depersonalization

Seth, Suzuki and Critchley (2011) were perhaps the first to apply predictive processing to depersonalization, and, since then, Gerrans (2019) has also provided an account. Seth et al. (2011) build their account on the central notion of "conscious presence." Since "presence" involves both a sense of oneself as present in the world, and the world as present to us, it casts depersonalization and derealization as two sides of the same coin. To briefly summarize their account, they build on work in schizophrenia research on the loss of the sense of agency (e.g. Frith 1987, Blakemore et al. 2000) according to which this arises from imprecise predictions about the sensory consequences of actions (see also our section 4.1, above). This account gets adapted to account for presence. According to Seth et al. (2011),

> "presence is the result of successful suppression by top-down predictions of informative interoceptive signals evoked (directly) by autonomic control signals and (indirectly) by bodily

> responses to afferent sensory signals. According to the model, disorders of presence (as in DPD) follow from pathologically imprecise interoceptive predictive signals". (p.2)

Our account builds on this in a number of respects. First, this account is based on a view of emotion as interoceptive inference. So is ours, in a sense, but what Seth and colleagues mean is emotion as interoceptive *perceptual* inference. In other words, as they explicitly state (p.1), they are fleshing out the James-Lange theory of emotion (James 1890) according to which emotion is perception of bodily (specifically visceral) change. Given a PP gloss, whereas perception is the result of model selection for minimizing prediction error from sense-perception, emotion is simply model selection for minimizing prediction error from interoception. This is perceptual inference since the model has to accommodate the input. Building on our recent work (Miller & Clark, 2017; Wilkinson et al. 2019), we, in contrast, view emotion, and affect more generally, as involving *active* inference too. In terms of ACM, it is a central part of allostasis. This brings us to another crucial difference with our view. The view of emotion as interoceptive model-building tells us nothing about *valence*. And yet, emotion has valence: it tends to be either positive or negative (and to greater or lesser degrees). In contrast, we tie positive valence to allostatic control (and negative valence to lack of such control). This means, crucially, given what we say later, that valence falls naturally out of our account, both of affect in general, but also of self-loss, both negative and positive, in particular.

Unlike Seth and colleagues' account, Gerrans' account doesn't take presence as a basic notion out of which both self and self-loss emerge for free. Instead, Gerrans appeals to the notion of a 'self-model', or, perhaps more accurately, the idea that the self-features as part of an overall predictive model that determines conscious experience. Gerrans' main point is that our ordinary experience of the world, and ourselves, is generated by a constant integration of cognitive, perceptual and affective signals. Building on his earlier work with Chris Letheby (Letheby & Gerrans 2017), the self here is part of a predictive model, one that works to explain away the affective changes that occur as the organism engages with its environment. When affective signals go missing the predictive system needs to explain the absence.

Gerrans' approach to explaining depersonalization focuses on the role of affect in the generation of our felt sense of presence (Seth, 2013). Like Seth et al. 2011, Gerrans concludes that when predictions about ordinary affective reactions are not fulfilled, the system generates the sense of the agent being no longer present in the experience. In short, Gerrans builds on Seth et al., but adds the self more explicitly into the model. To the extent that Gerrans' account is similar to Seth and colleagues' account, it shares many of the same differences with our own. Nevertheless, we like the embellishment of adding the self as a feature of the predictive model. In a sense, we would agree with Gerrans that presence emerges from a basic notion of self rather than the other way around.

What both of these existing accounts have in common, with respect to depersonalization, is the focus on affective numbing based on alterations (viz. inaccuracies) to interoceptive predictive processing. We don't disagree. However, what we add to the picture, is the idea that affect carries inbuilt valence, involves active inference (allostasis) and, crucially, plays a role in setting precision weighting. This has the welcome side-effect of allowing us to neatly explain other features of depersonalization beyond simply affective numbing. It also generates an account of depersonalization according to which it is inherently a negative experience (rather than something that needs to be appraised as such after the fact). These two other accounts tell us why there might be a loss of sense of self, but not why that is negative. And given the existence of extremely positive experiences of self-loss ('enlightened' states), the negativity of the experience is not something that should be taken for granted. Our explanation of this large difference in valence is that superficial similarities are masking quite radical differences in what is going on in the two cases.

In the next section we will propose that ACM can do a better job at accounting for depersonalization experiences than previous PP accounts. In particular, we will develop a view of depersonalization as a loss of allostatic control.

### 5.2. ACM Account of Depersonalization

Recall, the ACM casts the sense of self as underpinned by an inference about the system's endogenous control of self-evidencing outcomes. The highest levels of the self-model are also the most enduring, due to their being the most invariant across contexts. These higher levels that track how well we are able to control our interactions with the environment (i.e. allostatic control) more generally, act as hyper- priors informing more domain general precision estimations. The result of the sense of self being hierarchically deep in this way is that a temporary loss of control within a particular context may not necessarily reduce a more general sense of control, or a sense of control across contexts. In other words, someone can fail to play the violin well without losing confidence in their ability to live a good life.

This inference about allostatic control - manifesting as affective valence - plays a role in setting precision relative to changes in how well or poorly we are doing at reducing error given the context. Sensitivity to unexpected increases in prediction error rates - manifesting here phenomenologically as negative valence - acts as a disincentive to continue operating in a particular context (Kiverstein et al. 2017; Hesp et al. 2019). For example, when learning an instrument, if a certain song is too complex given our skill level the feelings of frustration that arise could motivate task switching perhaps to a simpler song, or to developing some of the skills necessary to eventually play the more complex tune. Task switching here offers a way for the system to get back to reducing error at a better rate.

But what happens when the system cannot resolve the negative affect through task switching? In other words, how would such a system behave if unexpected error continued to rise regardless of perceptual updates and behavioural interventions? For example, in active inference terms trauma could be understood as a massive influx of prediction error causing the system that it should drastically lower confidence (precision) in its action models (see Linson et al. 2020). In the case of physical trauma the body's integrity, which is highly expected, is seriously disrupted or damaged. The system in this situation is unable to reduce errors either by updating their models (perceptual inference) or acting in a way that will bring their expectations back in line with the current situation (active inference). This disparity between expected control of prediction error and the error-riddled reality produces huge amounts of negative affect, which as we have discussed above reduces certainty on the currently selected policy as a means of tuning the agent to better predictive opportunities.

If an external situation continues to create error (i.e. severe pain) over an extended period of time, and the agent cannot control the situation through switching domain or context (that would otherwise be controllable), the resulting drop in precision on expected free energy is going to be such that consequent transitions between higher level affective states will be forced into the same fearful state continuously. The ascending message from the negative 'affective charge' (Hesp et al. 2019) will override the descending message from higher-level policies (i.e. our ability to control error by task switching also fails to resolve the issue, and so we lose confidence domain general control). That means that you have exactly the same effect at the next level, whereby the desired state (positive affect) is never reached despite trying to control it, leading to a drop in the precision on expected free energy at that level, and so on upwards. Crucially though for this to happen, the person would have to never give up the resistance to the fear/pain. ie. maintain a high precision on that goal state (i.e. the phylogenetically expectation to have a healthy, well-functioning body). In time, the 'hopeless' situation might eventually create a learned belief that no matter what they do they will always be in a negative valence state (i.e. valence state transitions are not conditioned by policies and are stable in 'negative'). It would basically be a perfect storm for a gridlock situation where you have a strong preference against the negative state, but then also know you cannot escape it no matter what you do. The only option is to dissolve the process that is creating the negative affect in the first place since nothing else can work. The consequence would be an unravelling of the process by which we form affective states (i.e. inferring confidence in expected free energy), and with it the sense of self.[6]

The dampening of affect and the feeling of self-loss are intimately related here. In losing a sense of allostatic control, the system ceases to posit itself as a causally efficacious controller of sensations. Accordingly, in losing a sense that the system has allostatic control across contexts, the affective system ceases to tune to opportunities to reduce error. Nothing is motivationally salient because the system infers that it is not causally efficacious in bringing about self-evidencing outcomes, and as such it infers a global loss of confidence in precision on action policies (see Kiverstein et al. 2020). The result would be that the world would lose some of its *phenomenal depth* - it would cease to solicit one's engagements and so be perceived (just as DPD sufferers suggest) as two-dimensional, or flat (Medford et al. 2006: 93). The result is, as Colombetti and Ratcliffe write, that "The world ceases to matter, people and events are not salient anymore. With this, the world ceases to move and affect one through one's body" (2012: 148; see also work by Fuchs (2005) and Parnas & Sass (2003)).

This proposal casts new light on various circumstances of occurrence surrounding experiences of depersonalization. For example, consider a traumatic stressor such as torture, which is perhaps the most reliable instigator of depersonalization (Kira et al. 2013). On the current account, sustained inefficiency of the motivational system (in this case severe pain over an extended period of time) in a

[6] Thanks to Lars Sandved-Smith for discussions about the computational nature of depersonalization.

scenario where the person has no control to act to resolve the prediction error, results in a global loss of confidence in 'tuning' affective responses. Another example is major depression. Depression can be understood as a domain general inference of loss of allostatic control, where in extreme cases the system ceases to posit itself as a causally efficacious agent (Kiverstein et al. 2020). The chronic stress of an uncertain or volatile environment means eventually that the system ceases to posit endogenous control on the outcomes as the occurrence of positive/negative outcomes is inferred to be independent of the agent's actions. Through the current framework, the comorbidity of major depression with depersonalization can be understood to be a sustained loss of allostatic self-efficacy. Stephan and colleagues proposed that "the performance of interoceptive-allostatic circuitry is monitored by a metacognitive layer that updates beliefs about the brain's capacity to successfully regulate bodily states" (Stephan et al. 2016: 1), which they dub "allostatic self-efficacy". Other accounts have proposed that depression functions as a means of reducing prediction error associated with adverse social contexts (Badcock et al. 2017). If this is along the right lines, depression itself would function as a means of motivating withdrawal from potentially aversive contexts. The link between major depression and depersonalization can be understood here in terms of the system ceasing to posit itself as an endogenous controller of self-evidencing outcomes - when there are no possible context to move to (no high level, temporally deep goal priors to realise), the total loss of allostatic control is experienced as depersonalization.[7]

In the next section we turn our attention to an example of a potentially euphoric selfless experience, namely, the sorts of selfless experiences that can arise from meditation. A view of the self-model in terms of allostatic control gives a fitting account of this, and shows how this kind of selfless experience is radically different (indeed, relative to control they are diametrically opposed) to selfless experience in depersonalization.

## 6. ACM Account of Meditative Selflessness

While meditation is something of an umbrella term, the disciplined control of attention is central to almost all styles of meditation (Austin, 2013; Albahari, 2009; Garfield, 2015; Millière et al. 2018). For brevity, we will focus here specifically on *focused attention meditation*. Meditation here takes the form of

[7] Interestingly, this account is suggestive that the hyper-reflective tendency to check one's body and one's current state common in people suffering from depersonalization (Colombetti & Ratcliffe, 2012), could be understood as compensatory behaviour aimed at reducing the uncertainty relating to the perceived loss of control over interoceptive states.

consciously attending to a particular object (i.e. bodily sensation or breathing), and when the mind wanders from the chosen target and the practitioner realises the shift they actively "let go" of the distractor and reorient their attention back to the initial meditative target.

Active inference has recently begun to be applied to thinking about meditation (Pagnoni & Guareschi, 2018; Farb et al. 2015). Lutz and colleagues (2019) have provided the first account of *focused attention* meditation in terms of active inference. Focused attention meditation is described as having two interrelated aims: the pragmatic activity of regulating attention on a particular object (i.e. the breath sensations, an object); and the epistemological activity of increasing one's understanding of the nature of the meditative object and the various distractors, in particular recognizing their dynamic and impersonal nature (Lutz, 28).

In active inference terms, the pragmatic aspect of focused attention meditation requires that top-down directed precision enhance the behavioural policies associated with stable attention on an object. The challenge for the meditator then is to maintain this policy although multiple other policies may be simultaneously active, and acting as competitors for selection (Pezzulo & Cisek, 2016). In meditation this ongoing competition becomes simplified to include only the policy of maintaining attention on the meditation object, and all other competing policies attempting to divert attentional resources elsewhere (e.g. spontaneous memories, future planning, homeostatic concerns, mind wandering, etc.). Inaction during this process is considered crucial as it is the process of setting top-down precision on the sensory signals associated with the meditative object that allows this dialectic between focused attention and distraction to unfold, and to be consciously attended to. Lutz and colleagues suggest that this quality of "inaction" corresponds to the subjective experience of 'letting go' of the various distractions (p. 28).

Over time the meditator can learn to allow the various distracting thoughts and sensations to arise and pass without disturbing their concentration, that is without disrupting the meditation policy of focused attention. In part this occurs through learning to actively reduce precision on the distracting (negative) goal-prior. Inaction itself does this to some degree. An itch motivates, via negative valence, a scratching policy because of the preference for the non-itching state (i.e. goal-prior) and the high precision on the itching state (i.e. resistance to the sensations), as well as the learned connection between the itch and the action policy scratching. By not acting, and calmly observing the itch, the negative preference loses precision (i.e. resistance to the sensations is dropped). The result is that the probability of the sensation driving the selection of a new (distracted) policy is also lessened. In other words, meditators can actively reduce certainty on goal-states and mitigate involuntary policy selection through inaction. The result is that over time such goal-priors cease to draw processing (i.e. attention) in the same way.

In line with our view that the system infers its own control in terms of correspondence of action-outcome contingencies, here the system learns endogenous control on the precision of goal priors through repeated reduction via opting for the focused attention policy. Over time, the decrease in distractibility can be understood as an increased ability to endogenously control precision on goal priors. The system here refines attentional selection through mental action in a way analogous to refining motor commands through iterative inference of control of action-outcome contingencies (Miall & Wolpert, 1996; Wolpert & Flanagan, 2001). Note, that learning to actively adjust precision on goal-priors in this way is to learn to exert exactly the kind of control that is missing or weakened in cases of depersonalization as characterised above. In those cases, an inability to reduce precision on a goal prior that was unattainable leads to tremendous suffering and a loss of confidence in our ability to control the world more generally. In learning that it can endogenously set precision, it now begins to infer a domain general sense of being able to realise its goal states. This increase in domain general control may correspond to the pervasive feelings of joy and peace that characterize long term meditation (Dambrun, 2016).

During the course of becoming an adept meditator one develops the ability to remain poised between the object of attention and the ongoing flow of spontaneous mental activities. This capacity to remain subtly focused on the meditative object, while at the same time observing clearly the spontaneous mental activity, provides an optimal opportunity to learn about the process of policy selection taking place within its own system (Ridderinkhof et al. 2004). This is the epistemological element of focused meditation. The result of this introspective investigation is a gradual 'opacification' of these mental processes (Carter et al. 2005). A mental process is considered transparent insofar as its contents are available to consciousness, while its non-intentional structure or construction process are not (Metzinger, 2003). Without having access to the earlier stages of processing, transparent processes are presented subjectively as fundamentally real and personally essential. Metzinger writes,

> "Transparent phenomenal states make their representational content appear as irrevocably *real*, as something the existence of which you cannot doubt. Put more precisely, you may certainly be able cognitively to have doubts about its existence, but according to subjective experience this phenomenal content – the *awfulness* of pain, the fact that it is *your own* pain – is not something you can distance yourself from. The phenomenol-ogy of transparency is the phenomenology of direct realism and in the domain of self- representation it creates the phenomenology of identification..." (2017: 248).

By continually letting go, the distracting policy selections are observed as non-essential - they arise, they persist for some time, and they eventually dissolve without the need for overt actions. The system becomes aware of the constructed nature of these precision assignments (e.g. itching directly leading to scratching). Through repeated observation of this process the system ceases to identify with the precision selections, exactly because it observes them to occur without attributing them to a 'self' as an endogenous cause. Recall that the system infers itself as a self—as an endogenous cause of sensations—through agentive actions, where the correspondence between action-outcome contingency feels agentive. Through repeated reorientation of attention back to the meditation target (by re-engaging the meditation policy) the automatic precision assignments (e.g. itch → scratch) cease to be identified as essential, and so begin to lose the quality of immediateness and irrefutability that come with being transparent. Crucially, as the system observes these precision allocations occurring independently of agentive engagement (due to the non-action policy currently being engaged), the usual means of inferring itself as a self due to the correspondence of action-outcome contingencies is disrupted, and processes occurring in the system increasingly appear as nonessential to the self. This point about opacification is an important one for our discussion about selfless experiences.

This ties in closely with the 'non-self' themes in Buddhism. Common to all Buddhist schools is a critique of our ordinary self-experiences. The Buddhist doctrine of *no-self* (anattā) teaches that our ordinary self-experience is both mistaken and an important source of human suffering. The mistake is the common assumption that behind the various psychophysical processes that make up our conscious experiences, there is a single, essential subject, who is responsible for constructing and owning those processes. However, when we turn our attention inwards in an attempt to catch a glimpse of this assumed subject, all we ever experience is the dynamic and impersonal processes. This is the point - while we commonly assume there to be an essential and *unconstructed* subject, that sense of being a subject is in fact *constructed* from the interaction of our cognitive-behavioural processes. The Buddhist move here isn't to deny that the sense of self is real, that quality of experience that delineates my sensations from yours. Rather, the selfless insight Buddhist meditators seek out is the transformative recognition that while the self unreflectively appears to us as essential, persisting and unified it is in fact constructed, impermanent and dynamic (see Davies & Thompson, 2017).

According to the Buddhist tradition our mistaken assumptions about the self are generated and maintained by craving (*tanha* in Pali). Craving is a technical term here, it describes the felt urgency or motivational drive to make the world conform to our desires - our anxiety to perpetuate positive feelings and reducing negative ones. This ongoing emotional investment is what unifies the various impersonal psychophysical processes under a single idea: a persistent and essential self. In turn, the more we identify with our specific concerns or roles, the more intense the motivation to bring about those states in the world[8]. In humans, these desires expand beyond basic homeostatic concerns (i.e. being fed and watered) to include the wider constellation of ideas and roles we appropriate into our identity (i.e. being a student). In Buddhism, this craving is thought to be responsible for significant human suffering. Consider the difference in magnitude between the relatively short-lived pain of breaking an arm, and the potentially life long suffering of losing an opportunity to play a beloved sport professionally. At the heart of Buddhism is the teaching that while pain is unavoidable (i.e. the broken bone), our reaction to it (i.e. the craving to be a professional player) is optional. It is the craving, and not the pain, that is thought to be transformed through meditation, and with it the mistaken sense of self that craving engenders.

Meditation is presented as a vehicle for reducing craving and disrupting the mistaken view of self. In the satipatthana sutta, the foremost early Buddhist text on meditation, the student is directed to divide their experience into five categories (or "aggregates") of phenomena: bodily form, valence, perception, volitional activities and consciousness. These aggregates together are thought to create the experience of being a subject (Hamilton, 2000; Shulman, 2014). The aim of meditation here is to reduce one's identification with these psychophysical processes by closely examining each

[8] For a neuroscientific account of this relationship between identification and motivation see Damasio 2003.

individually and their interactions, and systematically noting their impersonal nature. In Anattalakkhaṇa Suttam the Buddha suggests that meditators observe each aggregate, saying to themselves 'not-I' or 'not-my-self'. As the various psychophysical processes that make up the self come to be experienced as 'not-I', the constructed nature of the self becomes apparent.

In addition to non-meditation policies (i.e. distractions) being driven by changes in goal-states, they are also driven by our affective reactions to those goal-states. As we have seen, valence sets precision relative to our ability to reduce error given a certain goal. Ordinarily, once a goal state is selected (predicted) any hesitancy in responding in the ways that the system has learned to expect produces negative valence, which has the effect of increasing the drive on policy selection as an attempt to catch up to the predicted slope. As long as valence is experienced transparently it has this powerful motivating effect on actions. The meditator then must also be able to reduce the certainty on the valenced reactions that occur from their commitment to non-action. By attending to the valenced sensations themselves (i.e. the discomfort of not reacting) the system learns that these signals too are changing and non-essential. As soon as one has made the observation that non-action makes them uncomfortable, these systems are already being rendered opaque. Focused attention meditation then simultaneously makes opaque precision on goal-priors (i.e. the confidence in the degree of wanting or not wanting certain states) and precision on expected free energy reduction that is given affectively.

In Buddhism, reflecting on valence (*vedana* in pali) is considered especially important in the process of relinquishing craving and disrupting the mistaken view of the self. Valence is considered the "weak link" in the process that gives rise to both craving and the mistaken view of self (Anālayo, 2009). In non-meditators, valence conditions craving: pleasant feelings give rise to attachment, painful feelings give rise to resistance. In contrast, long term meditators are thought to be able to experience valence without further craving-driven responses. The Buddha taught,

> "Touched by that pleasant feeling he does not lust after pleasure or continue to lust after pleasure. That pleasant feeling of his ceases. With the cessation of the pleasant feeling, painful feeling arises. Touched by that painful feeling, he does not sorrow, grieve, and lament, he does not weep beating his breast and becomes distraught" (Ñānamoli & Bodhi 1995:334).

Notice that even for the "well-taught noble disciple" (that is, meditators who no longer operate under craving and illusory notions of the self) valenced states still arise. These states are in and of themselves neutral in terms of wellbeing (Harris 2018). It is the strong motivational impulse to act (i.e. fleeing our pain; grasping at joys) in response to those signals that the Buddhist meditative project aims to transform. A close meditative investigation of valence is taught to have the effect of separating valence from craving. The result being that one begins to react with less preference relative to pleasure and pain. As Buddhist scholar Albahari writes, "As mental suffering is finally eliminated through insight [into the non-self nature of these processes], unpleasant *vedanā* will be confined to only physical (not mental) suffering" 2014: 11). As our urgency to react in self-serving ways diminishes, so too does the illusion of being an enduring and essential subject. The culmination of this process of extinguishing craving is *nibbāna* [enlightenment]: "the final flash of insight that burns out *taṇhā* and the sense of self for good" (ibid. p. 11).

The gradual opacification of valence results in various positive effects discussed in the Buddhist paradigm. As described above, inaction would allow for the opacification and dereification of the valence system. As valence is increasingly modeled (made opaque) it ceases to invoke that powerful sense of urgency (associated with the Buddhist notion of craving or taṇhā; Albahari, 2014), that, as we saw above, results from valence being experienced transparently and so presented phenomenologically as both immediately real and essential to the self (Metzinger, 2003). This change is an important one, as it results in the loss of the driving force that perpetuates the sense that there is an essential subject within and behind the various processes (Albahari, 2014). As valence is made opaque, and control over goal prior precisions is achieved, craving ceases due to its disentanglement from the conditioning influence of pleasure or pain. Thus, the illusion of self is disrupted. In gaining endogenous control on the precision of goal priors, meditation therefore enables the system to pull these apart so that (dis)liking doesn't need to condition (not) wanting, allowing the suffering associated with craving (and aversion) to be avoided. Gaining insight into the process by which valence is driven by our goal priors would have the consequence of allowing one to actively separate the two darts discussed by the Buddha. One comes to understand how pain (deviation from a goal-prior) leads to suffering (transparent negative valence signalling the deviation from the goal-prior and its expected resolution); that suffering occurs only insofar as we desire to avoid the pain (high precision on the goal prior drives error dynamics); and finally, although we often cannot do anything about the pain itself we can, by watching the whole process closely, render the valenced reactions opaque and so reduce the degree to which valence drives policy selection. That is, we can reduce the craving and suffering that arises from transparent valenced reactions simply by observing closely the link between pain and discomfort, thereby rendering the valence opaque.

The opacification of this part of the precision machinery opens new opportunities for control. Observing our valenced reactions allows us to develop new higher order policies about how precision (via valence) is being set on policies[9]. In other words, instead of valenced signals adjusting precision on policies directly, and so automatically conditioning us to behave in certain ways, one can now learn to activate alternative policies depending on the usefulness of the valenced signals. For example, mindfulness has been shown to be highly effective in helping people to quit smoking cigarettes (Bowen & Marlatt, 2009). The practice here is to disrupt the pattern leading from craving to using by selecting a policy to closely attend to the feelings of craving - the negative valence driving processing towards the expected rate of error reduction relative to nicotine levels - every time they arise. The new goal now activated every time there is a craving (instead of smoking a cigarette) is to watch as closely as possible the arising, the progression and the inevitable depletion of the craving-related feelings. Overtime, these sorts of mindfulness practices have the effect of teaching the system that cravings are in fact just feelings in the body, that can be allowed to direct processing and behavior or not.[10] This discovery can represent a major return of control for people struggling with substance addiction. Notice that valence here does not go missing through this process of opacification (as it does in depersonalization). Rather, valence begins to be interpreted by the system as what it is: information that can be useful, but is not essential, in selecting policies.

## 7. Conclusion

In this chapter, we seek to better understand the nature of the self from the perspective of the increasingly popular active inference framework. Within this framework, we present a novel account of the self, in terms of an allostatic control model (ACM; Deane, 2020). Central to the ACM account is a view of affect as a second-order process that guides the predictive system (via precision weighting) towards opportunities to improve. This is a novel take on affect that we have been

[9] See Sandved-Smith et al. (2020) for a computational model of this process.

[10] Notice here that the common phenomenology of long-term meditators being able to simply 'let go' of mental and emotional distractions is closely related to the *increase* in control that we are proposing this meditative process engenders.

developing over a number of recent publications (Kiverstein et al. 2017; Miller et al. 2020; Kiverstein et al. 2020).

We then put this account to work in trying to understand two starkly contrasting forms of self-loss, namely, depersonalization and the selfless experiences attained through meditation (viz. Buddhist no-self insights). Given our proposed framework, these two varieties of selfless experience are characterized by stark differences in the systems degree of control: whereas depersonalization is expressly characterized as resulting from a critical loss of inferred control, selflessness in the context of meditative practices is marked by a significant gain in control.

There is today, however, an increasingly popular idea that depersonalization and the selfless experiences attained through meditation are somehow closely related. Meditation teacher and neuroscience enthusiast Shinzen Young has called depersonalization the 'evil twin' of the Buddhist notion of *enlightenment* (Lofthouse, 2014). Part of what motivates this association, so it seems, are similarities in first personal accounts of both kinds of selfless states. Sufferers of depersonalization and long-term meditators make surprisingly similar reports about reductions in their experience of being agents of their actions, and as owners of their thoughts and behaviours. While these first personal accounts can sound very similar, given our framework this is where the family resemblance ends.

As we have shown throughout this chapter there are important computational differences between these two selfless experiences. Of particular importance, is the difference in how affective valence contributes to either state. Depersonalization is characterized as a loss of control, leading to the dampening of these affective valence systems. In meditation, there occurs a gradual opacification of the affective valence system. This opacification produces an important change in the meditator's relationship with the positive and negative affect - specifically, by no longer being automatically appropriated into their self-model. This has the result that changes in valence no longer create the existential urgency for change they would otherwise. It is important to note here that valence remains perfectly intact, continuing to tune the agent towards opportunities to improve in their predictive success. The purpose of the practice is not to disrupt valence itself (as happens in depersonalization), but rather to become conscious of the precision estimators in a way that allows them to select more skillful and beneficial policies. What is permanently altered in the meditation process is the system's reaction to those signals. While there is still the experience of frustration and joy there is no longer the sense of being an essential subject to appropriate these states as "me" and "mine". This insight leads to an eventual dissolving of our misguided idea that we are a single and enduring thing, to be replaced by an acknowledgment that we are a dynamic, self-organizing process. Far from reducing control, and in direct contrast to depersonalization, this development of one's metacognitive abilities here allows one to contextualize and control precision estimations in new and powerful ways.

Notice then, that for someone to transition from experiences of depersonalization to the selfless states attained through meditation they would first need to regain their phenomenal access to those affective responses. Without affect playing its role in tuning the system (relative to its predictive success) the opacification and subsequent insight into the nature of those precision processes could not occur. To be clear, we are not suggesting that the endeavour of bringing affect back online for people suffering depersonalization should be carried out through meditation specifically.[11] Rather, our point is that those affective signals would have to first become available for introspective access in order to be modelled in the way facilitated by focused meditation. This follows from the fact that, on our account, it is not the loss of affectivity that results in the selfless experiences sought after by Buddhist meditators, but the process of modelling those affective changes that opens the way for a new perspective of the self and a new layer of control to emerge.

In terms of control, depersonalization and selfless experiences sought after by mediators are, computationally speaking, polar opposites. And yet, there is something important to be said here about the potential focused attention meditation has to provoke depersonalization experiences. Britton's *Dark Night Project* has documented and investigated a large number of personal reports on various difficulties that can accompany meditative practices.[12] Britton (2019) suggests that adopting a persistent attitude of turning towards difficult stimuli and focusing on negative emotions can lead to negative outcomes. This aspect of mindfulness training has positive effects for some people, primarily by helping them to facilitate a gradual sensitization to negative affect. Such exposure approaches to therapy are thought to work by reducing avoidance, which has been shown to play a

[11] See Lindhal and Britton (2019) for reasons why that might be challenging.

[12] See also Segal (2002) and Lindahl & Britton (2019) for clear accounts of the relationship between meditation and depersonalisation.

leading role in the generation and maintenance of various psychological disorders (Barry et al. 2015). However, exposure therapies are most effective for people who tend towards high levels of avoidance (McNally, 2018). In other, low-avoidance personality types, anxiety and dissociative disorders can be produced and exacerbated by facilitating an attentional bias towards threat (MacLeod 2002; Eldar 2008). The fact is, the most effective treatment for an individual will depend on their baseline attitude towards threat. Given our account above, it makes sense why inappropriately meditating on traumatic events (that is beyond a certain healthy window of tolerance) could create depersonalization effects. If depersonalization is an airbag deployed when fight or flight won't work for getting one out of a traumatic emotional experience, then persistently meditating on an overwhelmingly traumatic experience while practicing inaction produces just those conditions. In effect, it could act like a kind of psychological self-torture. Meditation, in this case, would begin to lead to a perceived loss of allostatic self-efficacy, rather than towards the liberating states of self-understanding it is meant to.

# Chapter 5: Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution

Predictive processing approaches to brain function are increasingly delivering promise for illuminating the computational underpinnings of a wide range of phenomenological states. It remains unclear, however, whether predictive processing is equipped to accommodate a theory of consciousness itself. Furthermore, objectors have argued that without specification of the core computational mechanisms of consciousness, predictive processing is unable to inform the attribution of consciousness to other non-human (biological and artificial) systems. In this chapter, I argue that an account of consciousness in the predictive brain is within reach via recent accounts of phenomenal self-modelling in the active inference framework. The central claim here is that phenomenal consciousness is underpinned by ‘subjective valuation’—a deep inference about the precision or ‘predictability’ of self-evidencing (‘fitness-promoting') outcomes via action. Based on this account, I argue that this approach can critically inform the distribution of experience in other systems, paying particular attention to the complex sensory attenuation mechanisms associated with deep self-models. I then consider an objection to the account: several recent papers argue that theories of consciousness that invoke self-consciousness as constitutive or necessary for consciousness are undermined by states or traits of “selflessness”; in particular the “totally selfless” states of ego-dissolution occasioned by psychedelic drugs. Drawing on existing work—that accounts for psychedelic-induced ego-dissolution in the active inference framework—I argue that these states do not threaten to undermine an active inference theory of consciousness. Instead, these accounts corroborate the view that subjective valuation is the constitutive facet of experience, and they highlight the potential of psychedelic research to inform consciousness science and computational psychiatry.

## 1. Introduction

Phenomenal consciousness—the "what-it-is-like" (Nagel, 1974) to experience—has now been an area of serious scientific study for at least 30 years (Seth, 2018). Despite these efforts, a recent paper describes a "conceptual quagmire" in the science of consciousness (Doerig, Schurger, & Herzog, 2020) stemming from the propensity of theories of consciousness to identify a single process or mechanism underpinning conscious experience. Seth & Hohwy (2020), in response to Doerig et al (2020), argue that this results in these theories failing to deliver on a key desideratum of a theory of consciousness—namely, the ability to contrastively explain conscious phenomenology in terms of underlying mechanisms, as opposed to merely positing the presence or absence of consciousness in a given system. On their view, predictive processing is situated to do just this.

However, while predictive processing has been remarkably fecund in generating empirical predictions for consciousness science, constructing a theory of consciousness within predictive processing remains in question, largely due to the fact that predictive processing is not exclusively concerned with conscious processing (But see: Clark, 2019; Clark, Friston, & Wilkinson, 2019; Friston, 2018; Friston, Wiese, & Hobson, 2020; Hohwy, 2012; Hohwy & Seth, 2020; Kirchhoff & Kiverstein 2019; Ramstead, Wiese, Miller, & Friston, 2020; Safron, 2020; Solms, 2019, 2021; Solms & Friston, 2018; Whyte, 2019; Whyte & Smith, 2020; Wiese, 2018; Williford, Bennequin, Friston, & Rudrauf, 2018). Doerig et al (2020) emphasize that predictive processing, as it stands, is insufficiently constrained to stand as a theory of consciousness. On their account, predictive processing as a theory of consciousness is vulnerable to what they call the 'other systems argument'. A theory of consciousness "should be able to determine which systems, apart from awake humans, are conscious" (p. 7)—such that it can make clear-cut and specific predictions about which other systems are conscious. They contend that predictive processing fails to do this due to the fact that "there is no computational understanding of the crucial characteristics" (p. 21) that define the conscious condition within predictive processing. In this chapter I aim to show that predictive processing, in particular in the recent formulations of the active inference framework, has the resources to deliver a fully-fledged theory of consciousness.

Active inference is a process theory of the free energy principle (Andrews, 2020; Friston, 2010). Living systems, on this approach, can be understood to embody statistical models of their worlds, where they are biased towards the realization of 'phenotype-congruent' outcomes (Ramstead, Kirchhoff, & Friston, 2020). On this view, agents are "self-evidencing" in that they act in order to maximize the evidence for their own existence (Hohwy, 2016). The breadth of explanations within this approach—from microscale explanations applied to understand the adaptive behaviour of bacteria (Tschantz, Seth, & Buckley, 2020), and plants (Calvo & Friston, 2017), all the way up to social and cultural dynamics (Veissière, Constant, Ramstead, Friston, & Kirmayer, 2019) and natural selection (Campbell, 2016)—brings the need to identify the particular processes associated with consciousness itself into sharp relief.

The structure of this chapter is as follows. In section 1, I argue that consciousness arises as the system infers precision on control of self-evidencing outcomes across multiple levels of the hierarchical generative model. I then consider how subjectivity is 'shaped' by deep self-models, by considering some examples of disruptions in ordinary self-consciousness. In section 2, I explore how this characterization can inform attribution of consciousness to other systems, through identification of complex sensory attenuation mechanisms associated with consciousness on the current account. Section 3 considers an objection, the 'selflessness challenge'—that theories of consciousness that equate consciousness with self-consciousness ('subjectivity theories') are challenged by selfless experiences, most notably experiences of 'ego-dissolution' occasioned by psychedelic drugs (Letheby, 2020; Millière, 2020). In section 4 I build on an existing account of ego-dissolution in the active inference framework, (Deane, 2020), with a particular focus on the affective tone of the experience. In section 5, I respond to the selflessness challenge; I argue that understanding ego-dissolution in the active inference framework accounts for how the system can still be conscious without the typical structure of experience provided by deep self-models; and instead, these accounts in fact corroborate the view put forward in this chapter—that subjective valuation is the constitutive facet of experience.

## 2. Consciousness in active inference

In this section I argue that consciousness is underpinned by hierarchically deep self-models understood in terms of precision control in active inference. On this view, agency and phenomenal selfhood are inherent in active inference (Limanowski & Friston, 2020; Limanowski & Friston, 2018), as optimal action planning rests on the notion of *control*—where the system infers its control of sensation via action to realise self-evidencing outcomes.

The notion of a hierarchical generative model lies at the centre of this framework. A generative model is specified in terms of probabilistic beliefs about how observations relate to the states of the world that cause them (the likelihood), beliefs about how the states evolve over time; and prior beliefs—beliefs about the state of the world prior to observation. Inference here corresponds to inversion of the model in computing the probability of the unknown or hidden causes of the impinging sensory signals. As it is intractable to compute this posterior directly, approximate Bayesian inference is made tractable through optimisation of a posterior though variational inference—such that variational free energy is minimised. In this way, the approximate posterior converges towards the true (unknowable) posterior through the minimisation of variational free energy (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). Crucially, free energy can also be interpreted as a bound on the evidence for a generative model. This means that minimising the free energy just is maximising model evidence—hence the notion of self-evidencing (Hohwy, 2016; Palacios, Razi, Parr, Kirchhoff, & Friston, 2020). Self-evidencing is a technical notion that can broadly be understood as staying in a "species-specific window of viability" (Clark, 2013, p. 13).

For clarity about how consciousness can be conceived in the active inference framework, I unpack the inferential architecture of the hierarchical generative model and how this architecture relates to conscious contents in the following four subsections: i) basic perceptual models; ii) precision-weighting in perceptual inference; iii) planning as inference and the action model (Attias, 2003; Matthew Botvinick & Toussaint, 2012; Kaplan & Friston, 2018; Millidge, 2020); and iv) affective inference as precision on the action model—closely related to interoceptive and emotional inference (Fotopoulou & Tsakiris, 2017; Seth & Friston, 2016; Smith, Lane, Parr, & Friston, 2019).

## 3. Perception

A simple starting point to think about the generative model is to consider it in the case of moment-to-moment perception, which can be understood as state estimation or the brain's 'best guess' of the hidden causes of the incoming sensory signal. This basic (illustrative) generative model maps the relationships between observations (o)—the incoming sensory data, and the hidden states (s) in the world that caused the sensorium. A 'likelihood mapping'— encoding the probability of an observation under a generative model, given its causes (the hidden states in the world)—captures this relationship. In other words, what is the probability of the observations *given* the world is in a certain state? Formally, this is denoted as a likelihood model P(o|s). The full generative model includes beliefs about the most likely state of the world prior to any observation, which is known in Bayesian inference as the 'prior', denoted by P(s).

Of course, what the system needs to infer is the not the probability of observations given hidden states, but the inverse, the probability of hidden states given the observations. This is achieved through variational Bayes—movement from what the system has access to: observations, prior beliefs, and beliefs about how observations are caused by hidden states, to what it needs to infer: the hidden states that are the most probable causes of the incoming sensation—that is, the posterior probability of the states *given* the observations P(s|o). Inference about hidden states given observations is known as model inversion, as it is the inverse mapping from the consequences or outcomes to the causes. Model inversion finds the most plausible cause of observations, and as such, perception can be understood as 'posterior state estimation', i.e., estimating the hidden states and other variables that cause sensory outcomes. Formally, the process of updating a prior belief—on the basis of new sensory evidence—into a posterior belief is called Bayesian belief updating.

This basic machinery of perceptual inference has been applied to account for the perceptual contents of consciousness, such as the switching of perceptual contents in binocular rivalry (Hohwy, Roepstorff, & Friston, 2008). In binocular rivalry, subjects are presented with two different images to each eye—frequently a face to one eye and a house to the other. From the point of view of the subject, their conscious percept switches from one of the images to the other. According to Hohwy et al (2008), two competing models of what is present in the visual field are generated—one corresponding to a face, the other corresponding to the house. Because only one model can be accounted for at any particular time, the model corresponding to the non-dominant percept generates prediction error. Prediction errors are a central concept in predictive processing and the free energy principle. In a mathematically general sense, the free energy gradients that drive Bayesian belief updating in the free energy principle can always be expressed as a prediction error (in the form of a difference in log probabilities). In specific schemes, such as predictive coding, prediction errors are often treated as explicit variables that may be encoded by the activity of specific neuronal populations in the brain (e.g., superficial pyramidal cells). In the case of binocular rivalry, the prediction error generated by the non-dominant percept is thought to cascade up the cortical hierarchy until a critical threshold is reached, at which the previously dominant percept is suppressed—until the process repeats and the suppressed model generates enough prediction error to tip the balance back in its favour. In other words: "Over time, the prior probability of the currently assumed model (house or face, respectively) will decrease, leading to a revision of the hypothesis, until the brain settles into a state corresponding to the other percept, at least temporarily" (Wiese, 2014, p. 8).

## 4. Precision

Predictive processing architectures benefit from radical contextual flexibility afforded by 'precision-weighting'. Precision-weighting can be understood in terms of amplification or gain control, regulating the interaction between top-down and bottom-up signals by weighting them according to their expected 'precision'—where heavily weighted priors or prediction errors exert greater influence in determining the resulting posterior inference. Subjective precision can be understood as a prediction of the reliability of one's own beliefs, that is— for example, confidence in the likelihood mapping. Attentional processes are cast as implicitly metacognitive, in that they operate on second order statistics like precision: in other words, attentional states are beliefs about the (precision of the) system's beliefs. This means that precision—on the simple generative model just described—can be understood as the extent to which the system thinks observations reliably map to hidden states. Formally, precision is the inverse variance of a probability distribution (Feldman & Friston, 2010), and optimization of precision-weighting is frequently equated to attention within predictive processing schemes (Clark, 2013; Feldman & Friston, 2010). Heuristically, precision can be thought of as predictability or reliability of predictions—something that itself has to be predicted.

Predictive processing offers a picture of how a confluence of precision-weighted informational streams determines perceptual inference. Perceptual inference involves integrating information from across modalities to infer the hidden causes of sensation—for instance, during binocular rivalry, auditory (Lunghi et al, 2014), olfactory (Zhou, Jiang, He, & Chen, 2010), and tactile information (Lunghi & Morrone, 2013) have all been shown to influence which percept is dominant. Interoceptive (Salomon et al., 2016) and proprioceptive (Salomon, Lim, Herbelin, Hesselmann, & Blanke, 2013) information have been shown to affect visual experience using a continuous flash suppression paradigm.

Cue integration is one such example of how integration of (precision-weighted) informational streams give rise to the resultant percept. In a cue combination task, observers are presented with two more cues about a perceptual variable—such as, in early cue integration, the use of two depth cues (e.g., stereo and motion parallax) (Landy, Maloney, Johnston, & Young, 1995). The reliability of the cues can be varied (for instance by varying their visibility or contrast) to make one cue more reliable than the other. A series of studies have shown that when observers are asked to indicate their percept (a depth estimate), their estimates—depending on the cues, weighted by their precision, in combination with a prior probability—are approximately Bayes optimal (Ernst & Banks, 2002; Knill & Pouget, 2004; Landy et al., 1995).

Another example is gist perception in object recognition—where the 'gist' of the scene engages past experience to generate the most likely prediction about the object's identity (Bar, 2003; Oliva & Torralba, 2006). For instance, in the case of an ambiguous object, the context of a scene can determine whether the ambiguous input is perceived as a hairdryer or a drill depending on whether the context is in a bathroom or a workshop. These predictions are fed back to early visual areas to speed perception by constraining the hypothesis space of possible interpretations.

'Predictive penetration'—social, cognitive, and emotional—have all been demonstrated (O'Callaghan, Kveraga, Shine, Adams, & Bar, 2016). Importantly, precision is thought to mediate engagement with affordances—latent possibilities for action (Cisek, 2011; Pezzulo & Cisek, 2019); and sensory attenuation—the top-down filtering out of afferent (incoming) information, both from the body (interoception) and the senses (exteroception). As we will see, striking the right balance of precision in perceptual inference requires deep self-models.

The basic predictive processing story so far casts the brain as a hierarchical prediction machine using belief updating schemes to approximate Bayesian inference by utilising 'priors' (probability distributions about hidden states of the world), and incoming sensory data ('prediction errors') to arrive at a posterior estimate: a 'best guess' of the hidden causes of sensory signals (Aitchison & Lengyel, 2017; Clark, 2013, 2015; Hohwy, 2013). Predictive processing delivers a compelling story about the contents of perception, where "conscious perception is determined by the prediction or hypothesis with the highest overall posterior probability—which is overall best at minimizing prediction error" (Hohwy, 2012, p. 4). As yet, however, it is not clear why it should *feel* like something to perceive. In the next two sections, we see how these same principles can be built up to give an account of a *subject* of experience.

## 5. Active inference and the phenomenal self

To stay alive, organisms must maintain homeostasis, an 'internal balance' (Cannon, 1929), by keeping physiological states—'essential variables' (Ashby, 2013)—within reasonable bounds. The principles underpinning homeostasis have long been cast within the language of control theory (Conant & Ross Ashby, 1970), where homeostasis is achieved through autonomic control loops, such as sweating to lower body temperature. Creatures like human beings exhibit a form of prospective control or predictive regulation termed allostasis. In other words, they regulate the internal milieu by anticipating physiological needs and acting to meet them before they arise (Corcoran & Hohwy, 2017; Corcoran, Pezzulo, & Hohwy, 2020; Pezzulo, Rigoli, & Friston, 2015; Schulkin & Sterling, 2019; Sterling, 2012).

Active inference formalizes allostasis in terms of a single imperative—to minimize the divergence between expected and observed outcomes under a generative model that is fine-tuned over the course of phylogeny and ontogeny (Badcock, 2012). On the free energy principle, the basic imperative is to remain in 'expected' states—a species-specific window of viability. The free energy principle (Badcock, Friston, Ramstead, & Kauffman, 2019; Friston, Daunizeau, Kilner, & Kiebel, 2010), thus casts control theoretic 'essential variables' in terms of high precision prior expectations. This means that organisms are phylogenetically endowed with an expectation (and therefore bias to act) to, for example, maintain a body temperature within reasonable bounds. These high precision prior expectations are not amenable to perceptual revision, and instead must be fulfilled through corrective action (e.g., seeking out a shaded tree to maintain viable body temperature). Here, prior expectations are acquired through past experience, over the both the course of phylogeny and ontogeny (Badcock et al., 2019).

This means active inference is a formalization of allostasis as an inference problem; *planning as inference* (Kaplan & Friston, 2018) under the free energy principle (Friston, 2019). The planning and execution of action or a sequence of actions (a 'policy')—under this scheme—becomes a problem of inference, just like the approximate Bayesian inference of perception described earlier, but where the action with the highest prior probability is that which minimizes *expected* free energy—the expected dyshomeostatic consequences of an action policy (Kaplan & Friston, 2018).

## 6. Deep self-models in active inference

The phenomenal self-model can be understood as **“**the content of the conscious self: your current bodily sensations, your present emotional situation, plus all the contents of your phenomenally experienced cognitive processing” (Metzinger, 2004, p. 299). Phenomenal selfhood, is understood as being “The way you appear to yourself, subjectively, consciously.” (Metzinger, 2004, p. 26). Increasingly, the formal principles of self-modelling implied by active inference are thought to underpin phenomenal self-modelling (Deane, 2020; Deane, Miller, & Wilkinson, 2020; Friston, 2018; Hohwy & Michael, 2017; Limanowski & Blankenburg, 2013; Limanowski & Friston, 2018, 2020). On these views, “some notion of “self-hood” or “self-agency”—in the sense of inference about control—is inherent in active inference” (Limanowski & Friston, 2020, p2).

A central idea here is that, in acting, the system must infer *itself* as able to bring about the (self-evidencing) consequences of the action, where the self-evidencing consequences are understood in terms of expected free energy. Higher-order beliefs about *intentional selection* (‘What am I doing’) as opposed to beliefs about attentional selection (‘What am I seeing’) (Pezzulo & Cisek, 2019) are understood to underpin phenomenal selfhood on the current framework. As such, the self is seen as being a “hypothesis or latent state (of being) that can be associated with a self-model” (Limanowski & Friston, 2020, p3). In what follows, the inference about precision on intentional selection—cast in terms of an inference about ‘allostatic control’—is thought to underpin phenomenal selfhood.

Inference about the control of sensation via action—‘agentive control’ (Deane et al, 2020)—has been linked to the phenomenology of being an agent (Limanowksi & Friston, 2020). Agentive control is best understood as the system’s inference of its own ability to endogenously control sensory inputs via action (Hohwy & Michael, 2017), and as such is intimately related to the ‘sense of agency’—the experience of oneself as an agent who can cause events by acting (Haggard, 2017). Agentive control is understood here to be temporally deep, because expectations of the consequences of actions are not confined to the immediate future, but can predict abstract and distal outcomes (Pezzulo et al., 2015).

For example, an agent may expect proximal sensory consequences of tipping a watering can to water a plant pot, but also have temporally deep—and *abstract* (Gilead, Trope, & Liberman, 2019)—expectations about the form of the plant over the timescale of weeks and months. Recall, under active inference, lower-levels of the hierarchy track regularities that are unfolding on shorter timescales, and higher-levels track regularities unfolding on longer timescales (Friston, Rosch, Parr, Price, & Bowman, 2017; Kiebel, Daunizeau, & Friston, 2008). In just the same way that an organism can infer its own ability to control the immediate sensory consequences of action, by tracking regularities over time it can track its control of sensory outcomes more generally, where expectations of the downstream consequences of action inform policy selection (Friston, 2018).

Inference about agentive control is intimately related to the allocation of precision, and most specifically lowering precision to attenuate sensory evidence. There are two forms of sensory attenuation—'physiological' and 'perceptual' sensory attenuation—that critically relate to agentive control on this account (Palmer, Davare, & Kilner, 2016). The first to consider is *physiological sensory attenuation* (Palmer et al, 2016). Physiological sensory attenuation is critical for movement initiation (Brown, Adams, Parees, Edwards, & Friston, 2013). Action initiation involves systematic misrepresentation—whereby proprioceptive evidence that, for instance, my arm isn't moving, is attenuated to allow the system to bring about the desired movement (Adams, Shipp, & Friston, 2013; Brown et al., 2013). Higher-level prior beliefs attenuate current sensory evidence and higher precision is afforded the anticipated sensory consequences of the desired action. Prediction error is then suppressed by making the prediction come true, through reflex arcs at the lowest level of the hierarchy (Parr, Rees, & Friston, 2018). In the setting of motor control, this perspective on action is closely related to idea motor theory (Limanowski, 2017) and 20th-century formulations in terms of the equilibrium point hypothesis (Feldman & Levin, 1995). In other words, all that is required for intentional movement is a specification of the desired sensorimotor endpoint of a movement—and motor reflexes bring the motor plant to that equilibrium or setpoint. Another perspective on this formulation is perceptual control theory (Mansell, 2011), where action is in the game of bringing about desired sensory consequences—in this instance proprioceptive sensations from the musculoskeletal system. In order to move, then, the system predicts *itself* in the desired state. Physiological sensory attenuation aids in entertaining counterfactual hypotheses about oneself (Limanowski & Friston, 2020*)* in order to generate the self-fulfilling prophecy of moving. This self-attenuation needs only be applied transiently for movement initiation, but these same sensory attenuation mechanisms have been are argued to underpin various states of altered self-experience. For instance, in the "rubber hand illusion" (Botvinick & Cohen, 1998), visual information about the location of the hand is deemed to be precise due to the corroborating synchronous stroking pattern on the rubber hand, while the conflicting proprioceptive input suggestive of the hand's real location is down-weighted in order to maintain a coherent bodily representation (Limanowski & Friston, 2020).

*Perceptual sensory attenuation* is the top-down filtering of afferent information to limit how much feedback is received from self-generated movement. On the current account, perceptual sensory attenuation is critical to the formation of the on-going inference about agentive control. Originally developed as a theory of motor control, the 'comparator model' posits that motor commands are refined through comparing sensory consequences of an action with the intended consequences of an action (Miall & Wolpert, 1996; Wolpert & Flanagan, 2001). Subsequently, the comparator model has been used to account for the sense of agency (David, Newen, & Vogeley, 2008; Feinberg, 1978; Frith, 2005), where inference about endogenous control over the causes of sensory signals is thought to underpin the sense that an action is agentive or self-generated. For instance, sense of agency would be low in the case of a mismatch between motor output and sensory input, such as (when wearing a VR headset) a virtual hand that moved in a way that did not correspond to movements of the subject's real hand. The mismatch between the expected and actual consequences of a given action justifies the attribution of sensory outcomes to exogenous (external) rather than endogenous (internal) causes, such that attribution of sensory outcomes to exogenous causes results in a reduced or absent sense of agency (Sirigu, Daprati, Pradat-Diehl, Franck, & Jeannerod, 1999). Indeed, incongruent action-outcomes have been linked to a reduced sense of agency (O'Sullivan et al., 2018). A self-other distinction critically relies on balancing this attribution to exogenous and endogenous causes.

Selectively attenuating precision on sensory inputs allows the system to filter out irrelevant inputs (Crapse & Sommer, 2008a), such as those caused by self-generated actions. One such example of this is saccadic suppression, where, despite saccadic eye movements, perception of the environment remains stable. Reduced precision on afferent inputs from self-produced tactile sensation is thought to cause inability to tickle oneself (Blakemore, Wolpert, & Frith, 2000). Sensory attenuation in relation to movement and self-experience will be discussed in more detail in the sections on disturbances of self-consciousness and other minds.

## 7. Agentive control and self-evidencing

Inference about agentive control is concerned with control of self-evidencing outcomes. It is necessary for a system to infer agentive control because selection of optimal action policies involves having counterfactually rich expectations of the states of affairs that would be brought about contingent on actions (Friston et al., 2017; Pezzulo, 2017; Seth, 2014). In other words, to select action policies that maximize the self-evidencing outcomes over time, organisms rely on deep temporal models (Friston et al., 2017), that encode expectations about the evolution of states of affairs over time contingent on action policies, such that the system can infer actions that result in sensory states conducive to continued existence—sometimes called the "attracting set" (Friston, 2012)

Just as with perceptual inference (belief updating or 'state estimation'), action selection is similarly understood in terms of Bayesian model selection, where possible action policies are scored with respect to the expected free energy associated with pursuing a given policy. Here, the agent is equipped with beliefs about state transitions, where beliefs about states transitions are updated in light of the action or action policy that is currently being pursued. Conditioning state transitions on actions—in the generative model—allows the agent to select action policies that have the least expected free energy, where expected free energy can be decomposed into *epistemic* and *pragmatic* value, such that the agent can learn about its environment while realizing prior preferences (Friston et al., 2017, 2015). As the quantities that agents seek to control with action, these quantities are crucial in the construction of conscious experience—so it is worth unpacking each in turn.

### Pragmatic value and prior preferences

Agents are not disinterestedly inferring their control of sensation via action. Rather, in active inference the agent acts as a "crooked scientist" (Bruineberg, Kiverstein, & Rietveld, 2016), acting to realise prior preferences—changing the world to make it conform to prior expectations, as opposed to changing expectations to conform to the world (i.e., perceptual inference). Active inference thus recasts "essential variables"—physiological quantities that must remain within specific bounds for an organism to stay alive (Ashby, 2013)—as high precision 'prior preferences'. Prior preferences are phenotype specific states that the organism expects itself to be in to be in—connecting control to states of the body and views of selfhood based in interoceptive inference (Barrett, 2017; Seth & Friston, 2016; Seth & Tsakiris, 2018). Prior preferences about essential variables encode probability distributions over states (rather than a single ideal setpoint), and the sufficient statistics that specify this setpoint (mean and precision) are free to vary and can be toggled according to the context (Ainley, Apps, Fotopoulou, & Tsakiris, 2016).

This is key in allostasis, as it allows for temporary deviations from a homeostatic setpoint in order to realize sensory states in the "attracting set" on longer timescales. For instance, heart rate and blood pressure are more flexible to contextual alteration in order to realize certain actions (e.g., fleeing from a predator), while others such as blood pH and core body temperature may be less variable due to more constant high precision (Corcoran et al, 2020). While many prior preferences are phylogenetically endowed, over the course of ontogeny an organism will acquire prior preferences that subtend increasingly deep temporal scales. The expected free energy of a given policy, then, is going to depend to some degree on how much the given policy fulfils prior preferences, and so, as we will see, a critical part of the phenomenal self-model is understood in terms of an inference about control of the realization of prior preferences.

### Epistemic Value

Self-evidencing agents not only act in order to realize prior preferences, but they also engage in novelty seeking behaviours that realize epistemic value (Friston, Pezzulo, Cartoni, & Rigoli, 2016; Friston et al., 2015)**.** The epistemic value or affordance of a given policy refers to the information gain or resolution of uncertainty about the causes of sensation. Optimal epistemic action, or 'epistemic foraging', requires the agent to have beliefs about their own uncertainty, enabling action directed towards higher sensory precision. Agents minimizing expected free energy seek out observations that resolve ambiguity about the state and causal structure of the world. Curiosity and novelty seeking behaviour are accounted for within this formulation of epistemic action (Friston et al., 2015; Kiverstein, Rietveld, & Miller, 2017; Mirza, Adams, Mathys, & Friston, 2018; Pezzulo & Nolfi, 2019). This can be understood in terms of sensitivity to long-term epistemic affordances (Bruineberg & Rietveld, 2014; Parr & Friston, 2017). When the agent is confident about its model of the world, and epistemic value is much the same across policies, pragmatic or instrumental value (fulfilment of prior preferences) dominates behaviour.

## 8. Affective inference and phenomenal consciousness

In Thomas Nagel's paper *What is it like to be a bat?* Nagel argues:

> "An organism has conscious mental states if and only if there is something it is like to *be* that organism—something it is like *for* the organism." (Nagel, 1974, p. 436)

Why, then, in the active inference framework, is experience *felt?* The claim here is that phenomenal consciousness is underpinned by estimation of the precision of its own action model. To be more specific: the system needs to engage in subjective valuation, that is—set precision on competing action policies across multiple levels of the hierarchy based on inference about endogenous control of self-evidencing outcomes. Precision on action policies on this account is understood to be a fundamentally *affective inference.* The meaning of incoming sensory data *for me,* is understood here, as "*What does this mean for precision on my action model?".* As a confidence estimate in the action model, it is 'subjective' in the sense that it can be out of step with reality—i.e. the system could be over-confident or under-confident in these estimations. In order to account for the felt aspect of experience, this section will apply the same inferential machinery already described previously to account for affectivity (Hesp et al, 2021).

Affective inference—in terms of a contextually flexible inference of the precision on prior preferences and epistemic affordances—acts to "tune" the organism to possibilities for self-evidencing action in the environment. Precision on prior preferences is inferred across the control hierarchy (or 'deep goal hierarchy' Pezzulo & Cisek, 2019). Pain perception is a great example of 'tuning' affective inference. Precision is allocated to, for instance, the "healthy body condition" prior preference (Ongaro & Kaptchuk, 2019) according to a host of contextual factors. This flexibility enables organisms to "tune their own pain perception according to both their prior beliefs and the specific biological goals they believe are attainable in that context" (Moutoussis, Fearon, El-Deredy, Dolan, & Friston, 2014, p. 70). Mounting evidence speaks against the more classical view of pain as tracking tissue damage, in favour of a view of pain perception as underpinned by a process of inference. In particular, Bayesian models of pain perception provide evidence that affectively charged percepts are inferential in nature (Anchisi & Zanon, 2015; Morton, El-Deredy, Watson, & Jones, 2010). For example, studies show that patients who receive treatment in a medical context experience considerably higher pain relief than those who receive analgesic drug treatment covertly (Benedetti, Carlino, & Pollo, 2011; Benedetti et al., 2003). The felt intensity of pain can be adjusted according to the context and the survival needs of the animal, modulated by attention, expectation, conditioned pain modulation, and placebo responses (Atlas, Lindquist, Bolger, & Wager, 2014; Atlas & Wager, 2012; Kirsch et al., 2014; Kong & Benedetti, 2014). Even social information can have a profound influence on experience: other people's pain reports affected participants' pain experience and physiological indicators of increased pain such as the skin conductance response (Koban & Wager, 2016).

Inference about endogenous control of self-evidencing can thus be understood as an inference about "subjective fitness"— the expected precision of the organism's phenotype-congruent action model (Hesp et al, 2021). On this account, interoceptively registered bodily changes track how well the organism is doing at minimizing expected free energy—i.e., fulfilling prior preferences and resolving uncertainty (Joffily & Coricelli, 2013; Kiverstein, Miller, & Rietveld, 2020; Kiverstein et al., 2017; Seth & Friston, 2016). This contextually flexible evaluation of model fitness is essential for organisms to persist and perform adaptive actions in volatile environments. Promoting self-evidencing outcomes on longer timescales requires organisms to be sensitive not only to prediction error reduction in the present, but the rate of prediction error reduction over time (Joffily & Coricelli, 2013; Kiverstein et al., 2017; Van de Cruys, 2017). On this view, certain rates of prediction error over time—such as progress towards a goal—becomes a prior preference fulfilled by (temporally extended) action. As such, deviation from the prior preference manifests to the system affectively, acting as motivation to realize the prior preference via action. The roots of these approaches can be traced to control theoretic precursors that postulate a second feedback system that senses and regulates the rate of the action guiding system (Carver & Scheier, 1990).

Inference about the reliability of the action model allows the system to increase or decrease precision on the current policy (Hesp et al., 2021; Kiverstein et al., 2017). For instance, if the current policy is reducing prediction error at a rate that is worse than expected, this manifests to the system as negative affect, and acts as an incentive to discontinue the current course of action. Affective valence here is being reimagined within the active inference framework as a 'domain general controller' (Deane et al., 2020; Ramstead, Wiese, et al., 2020). Inference about how well the system can expect to reduce error via action *in general* is informative as it informs precision on policies across contexts, acting as a domain general prior on the precision of policies generated by the action model (Hesp et al., 2021, 2020).

Deane et al (2020) suggest that a sensitivity to worse than expected rates of prediction error reduction over time (Hesp et al., 2021; Kiverstein et al., 2020, 2017), manifesting phenomenologically as negative affect, drives the system to switch to more tractable goals. For instance, while loss of control in a particular context (such as learning to play a particularly difficult piece in a piece of music) might create negative affect, this negative affect functions as an incentive to switch to a task with a better expected rate of prediction error reduction. As such, loss of control in a particular domain does not necessarily impact a more domain general sense of control, related to more fundamental and pervasive sense of self as a causally efficacious agent. As such, affective inference—inferring precision on prior preferences and epistemic affordances across multiple hierarchical levels—tunes the organism to adaptive actions in the given context.

The preceding paragraphs give a picture of how the mechanisms underpinning phenomenal consciousness—a 'deep control model'—act to 'tune' the organism to adaptive action in the world. On this view, our status as 'beast machines' shapes our subjective experience (Seth & Tsakiris, 2018). Empirical evidence attests to this picture—for instance, it has been demonstrated that neutral stimuli are more often perceived as fearful when subjects were given (false) feedback of increased heart rate (Anderson, Siegel, White, & Barrett, 2012). Hierarchically (and temporally) deep contextualization of interoceptive signals tunes an organism to appropriate action and engagement with environmental affordances (Pezzulo & Cisek, 2019), and assigns appropriate weight to priors and ascending prediction errors across the cortical hierarchy. Notice that this means that even state estimation associated with perceptual inference is determined by the overarching inference about control of self-evidencing outcomes—both in terms of the predictive models encoding sensorimotor relations ("counterfactual richness") grounding the subjective reality of perceptual contents (Seth, 2014), and in terms of those perceptual contents being filtered through deep goal hierarchies (Pezzulo et al., 2015; Pezzulo, Rigoli, & Friston, 2018). As such, the conscious agent encounters "a structured world apt for action and intervention, and inflected at every level, by an interoceptively-mediated sense of mattering, reflecting 'how things are for me as an embodied agent'" (Clark, 2019, p. 7). This means that experience of the world is suffused with our "cares and concerns" (Ramstead, Wiese, et al., 2020), and accords with the view that visual perceptual experience is determined by the agent's ‘poise’ over the ‘action space’, where we encounter the world as a “matrix of possibilities for pursuing and accomplishing one’s intentional actions, goals and projects” (Ward, Roberts, & Clark, 2011, p. 1).

Precision on control at different levels of the hierarchy crucially allows the system to arbitrate between competing affordances on different timescales (Pezzulo & Cisek, 2019), providing the system with a common motivational currency for navigating trade-offs on different timescales. Conceiving of valence as a ‘common currency’ to arbitrate between action plans in this way connects this proposal to numerous accounts of phenomenal consciousness in the literature (Cabanac, 1992; Merker, 2007; Morsella, 2005). Moreover, contextual modulation of the precision on expected free energy is critically related to flexible behavioural control (Pezzulo et al, 2015), and as such bridges the current story to the association between consciousness and flexible behaviour (Dehaene et al, 2017).

## 9. Shaping Subjectivity: Disruptions in (self-)consciousness

Altered self-experience provides some of the most compelling illustrations as to how subjectivity is shaped through an inference about allostatic control. To illustrate this, this section briefly considers depersonalisation and meditation.

A domain general loss of precision control has been used to understand depersonalisation disorder (Deane et al, 2020). This account—through connecting views of affectivity in terms of precision estimation on expected free energy to the feeling of being an agent—casts the computational mechanisms of depersonalisation as an inferred loss of allostatic control, whereby the system ceases to posit itself as causally efficacious at realizing self-evidencing outcomes. As we saw in the previous section, the affective system usually acts to tune the system to action opportunities across multiple interlocking timescales. Depersonalisation is understood as occurring due to a global loss of precision on action policies, and as such the world loses “phenomenal depth”—as described by sufferers of depersonalization—in that it ceases to solicit engagement and is perceived as flat or two-dimensional (Medford et al., 2006). Major depression, similarly, has been characterized in terms “domain general inference of a loss of allostatic control” (Ramstead, Wiese, et al., 2020).

This phenomenology is contrasted with a perceived gain in allostatic control in meditation practitioners, as precision (on prior preferences, for instance) becomes increasingly under endogenous control (Deane et al, 2020). This account makes use of the fact that mental action in active inference follows just the same principles as the account of action initiation put forward earlier, but where the hidden states are *attentional* states (precisions), and the state transitions are transitions between attentional states (Smith et al, 2020). The idea here is that *focused attention meditation* (Lutz, Mattout, & Pagnoni, 2019) can be understood as the endogenous withdrawal of precision from prior preferences, due to the practice of repeatedly bringing attention back to the attentional object, such as the breath. For instance, the sensation of an itch can be understood in terms of increased precision on a scratching policy. Through withdrawing precision from the sensation and back to the target sensation the system learns an extra level of agentive control, that is—endogenous control of precision on prior preferences. Over time, this becomes domain general, such that the system learns to have precision control over its own affective system.

## 10. Consciousness in other systems

Let us return briefly to the question of whether this specification of the conscious condition within predictive processing and active inference can make predictive processing less vulnerable to the 'other systems argument'. (Ramstead, Wiese, et al., 2020) state: "only higher forms of life may have sufficiently deep or elaborated generative models to support this kind of affective or emotional inference". Inference about confidence in endogenous self-evidencing capacity, or precision on expected free energy—understood to underpin phenomenal consciousness in this chapter—crucially determines the allocation of precision on sensory evidence. This section sketches how the complex sensory attenuations mechanisms associated with consciousness on the present account can give clues as to the neuroanatomical substrates and processes possessed across species that are indicative of conscious experience.

Holst & Mittelstaedt (1950) identified an interpretative problem as to whether sensory signals arise from the environment or the animal's own muscles and movement, dubbed the 'reafference problem'. The reafference problem arises due to the fact that sensory receptors are indifferent to the cause of their activation, whether it be from exafference—occurrences in the environment, or reafference—inputs that result from an animal's own movements (Holst & Mittelstaedt, 1950). Sensory neurons are able to respond with high sensitivity to exafferent inputs despite disruptive self-generated inputs (Ahrens et al., 2012; Bell, 1981; Eliades & Wang, 2008; Keller & Hahnloser, 2009;

Poulet & Hedwig, 2002). Across species, the sophisticated filtration process underpinning this high sensitivity to exafferent inputs is thought to be achieved through the mechanisms of *corollary discharge*—predictions of the sensory consequences of actions that act to suppress reafferent inputs (Crapse & Sommer, 2008a). In predictive processing, corollary discharge can be understood simply as top-down predictions that explain away sensory prediction error (Friston et al., 2010).

Crapse & Sommer (2008) make a distinction between lower-order (reflex-inhibition and sensory filtration) and higher-order (sensory analysis and sensorimotor learning/planning) corollary discharge based on their underlying neuroanatomical substrates. Lower-order corollary discharge enable reflex inhibition and sensory filtration, by intervening so as to regulate and control sensation entering the central nervous system, and appear to have the function of "transient, protective inhibition of sensory networks" (Crapse & Sommer, 2008, p592). For example, the nematode C. elegans —often used to study simple nervous systems—has a simple behavioural repertoire and only 302 neurons and uses lower-order corollary discharge in order to inhibit reflexes that would be triggered by reafference. Barron & Klein (2016) argue that in this very simple nervous system —with only two layers separating sensory neurons from motor neurons—there is no evidence that this sensory attenuation mechanism contributes to a structured model of the self or a model of action-outcome contingencies informing selection from a range of possible actions. This is behaviourally as well as neuroanatomically apparent: when hungry, nematodes respond with increased locomotion in a random search pattern (Artyukhin, Yim, Cheong, & Avery, 2015; Lüersen, Faust, Gottschling, & Döring, 2014). By contrast, hungry rodents, ants and bees will direct their search towards locations where they have encountered food previously (Oades & Isaacson, 1978; Seeley, 2009; Wehner, 2013). In the case of the nematode, the corollary discharge does not seem indicative of a model of temporally deep control, and the lack of anticipatory and goal-driven behaviour makes it unlikely nematodes have phenomenal consciousness on the present proposal.

Crapse & Sommer (2008) identify higher-order corollary discharge as involved in predictive control in perceptual cohesion and action sequencing—this *does* seem suggestive of a deep control model. For example, bats explore their environment by emitting beams of sound and then comparing the emission with the spatiotemporal aspects of the returning echo and to construct a cohesive and counterfactually rich world-model. This complex process involves having predictions about regularities tracking multiple timescales (Kiebel et al., 2008), and the differences between the corollary discharge and the input are used to infer properties such as the size, speed and location of the object reflecting the sound. In action sequencing, higher-order corollary discharge is also involved in temporally extended planning strategies —for example, primates use corollary discharge to keep an internal record of the current saccade to facilitate planning the next saccade (Crapse & Sommer, 2008b). Complex reafferent processing is also shown in juvenile songbirds as they imitate the song of tutor (Brainard & Doupe, 2000; Margoliash, 2002). This requires refining on-going action plans via continuous updating of an internal record of current state, allowing for flexible contextual interpretation of sensory input towards the realization of temporally deep goals (Crapse & Sommer, 2008a).

These complex and context sensitive sensory filtration mechanisms are of the most promising places to look for hallmarks of consciousness in non-human animals. Peter Godrey-Smith—in considering the evolution of subjectivity—reaches a similar conclusion: "once animals start to accommodate and utilize reafference, the character of sensing changes. The animal is now not only open to the world, but open to the world as the world, as distinct from self." (Godfrey-Smith, 2019, p. 13). The deep control model of consciousness put forward in this chapter specifies why these mechanisms may be associated with subjectivity.

## 11. The selflessness challenge

Understanding consciousness in terms of self-consciousness and self-modelling aligns the current account with many other approaches across psychology, neuroscience, and philosophy that cast self-consciousness as necessary or constitutive of consciousness itself (Damasio, 1999; Gallagher, 2010, 2013; Lou, Changeux, & Rosenstand, 2017; Metzinger, 2004; Millière, 2017; Zahavi, 2014). For instance, variations on this claim have been made in the phenomenological tradition date back at least to Husserl, and more recently Dan Zahavi (2014) says that "[S]elf-consciousness is an integral and constitutive feature of phenomenal consciousness […]" (p. 62). Antonio Damasio (1999) argues "If 'self-consciousness' is taken to mean 'consciousness with a sense of self', then all human consciousness is necessarily covered by the term—there is just no other kind of consciousness as far as I can see" (p. 19). All these theories take on the idea that consciousness involves a kind of phenomenological centredness on the self as the experiencing subject, where consciousness entails a kind of self-consciousness.

This claim is embedded into the active inference framework. Friston (2018) states "Is self-consciousness necessary for consciousness? The answer is yes. So, there you have it—the answer is yes." (p. 1) On one hand, this appears to be an explanatory advantage of the active inference approach to self-modelling and consciousness—not only are the deep links between consciousness and self-consciousness formalized, but it provides theoretical underpinning for of host of closely related phenomena, including selfhood, emotion, attention, and the sense of agency.

On the other hand, however, several recent papers have argued that experiences of altered selfhood present a problem for theories of consciousness that claim self-consciousness is necessary for consciousness (Billon & Kriegel, 2016; Letheby, 2020; Millière, 2020). Millière & Metzinger (2020) highlight the fact that this view of self-consciousness as embedded into the very structure of experience may be what Dennett calls Philosopher's Syndrome: "mistaking a failure of imagination for an insight into necessity" (Dennett, 1993, p. 401). Billon & Kriegel (2016) take the cases of "inserted thoughts" in schizophrenia, and the disowned mental states of patients with depersonalization disorder, as problematic cases for proponents of the claim that self-consciousness is necessary for consciousness—as apparent cases where self-consciousness appears to be missing from consciousness. Arguably, however, as noted by Millière (2020), it is not clear these are the most difficult cases as these are likely only "partially selfless" states. Millière distinguishes between the 'necessity claim'— the claim self-consciousness is necessary for consciousness in general, and the 'typicality claim'— that self-consciousness is merely present in ordinary experience, and argues that the subjectivity theorist must be committed to the necessity claim. Millière distinguishes six different notions of self-consciousness that are commonly discussed in the literature, arguing that there is empirical evidence that there are states of consciousness where these states fail to be instantiated. These states of consciousness can be described as "partially selfless". Millière notes that none of the partially selfless states of consciousness would be sufficient to rule out a disjunctive version of the necessity claim—where any form of self-consciousness would be sufficient but not necessary for consciousness. However, there is evidence for states of consciousness that appear to be "totally selfless"—lacking in all the ways one could be self-conscious. Both Millière (2020) and Letheby (2020) take the "totally selfless" states of psychedelic-induced ego-dissolution to be evidence against the claim that self-consciousness is necessary for consciousness.

Serotonergic psychedelics such as LSD, psilocybin, DMT (found in ayahuasca), are known to produce profound alterations in phenomenology (Preller & Vollenweider, 2018). Most notably for present purposes, psychedelic experiences, especially at high doses, are characterised by profound alterations in self-consciousness (Huxley, 1952; Leary, Metzner, & Alpert, 1964; Lebedev et al., 2015). Both Millière (2020) and Letheby (2020) argue that the 'total' ego-dissolution induced by the serotonergic psychedelic 5-methoxy-N,N-dimethyltryptamine (5-MeO-DMT) is the strongest evidence against the view that self-consciousness is necessary or constitutive of consciousness. While there is empirical evidence that some advanced forms of meditation practice can also occasion 'totally' selfless states (Laukkonen & Slagter, 2020; Millière, Carhart-Harris, Roseman, Trautwein, & Berkovich-Ohana, 2018; Winter et al., 2020), here I will focus on psychedelics as the most robust catalysts of selfless states.

Consider these phenomenological reports of the 5-MeO-DMT experience retrieved from the database of drug experiences erowid.org, cited in Millière (2020) as evidence of "totally selfless" states:

> I was completely disassociated from the "real world" and [from] any sense of self. It was the most jarring feeling. (#107905)
>
> It is a complete annihilation of self […]. I was absolutely nothing but a sensory perceiver, stuck within the split seconds that were eternity. (#18198)
>
> It felt as if all of the atoms of the molecules that typically form my physical self simply dispersed, and even my sense of self, or ego, vanished […]. (#56384)
>
> I wasn't me any longer. There was no me. There was no ego. (#27601)

These experiences present considerable counterevidence to the necessity claim, due to being both vividly phenomenally conscious while totally lacking in any kind of ordinary self-consciousness. Do these experiences provide genuine evidence against the claim that self-consciousness is necessary for consciousness? More specifically, are they problematic for the active inference account of consciousness in terms of self-consciousness? To answer this question, the next section builds on

Deane (2020) to provide an account of ego-dissolution in active inference.

### Psychedelics and selflessness in active inference

The REBUS—"RElaxed Beliefs Under pSychedelics"—model casts the action of psychedelics in the predictive brain in terms of a 'relaxation' (lowering) of the precision of high-level priors, thereby liberating bottom-up information flow (Carhartt-Harris & Friston, 2020). Although a preliminary account, the REBUS model is evincing growing empirical support (Alamia, Timmermann, Nutt, VanRullen, & Carhart-Harris, 2020; Dupuis, 2020; Girn et al., 2020; Herzog et al., 2020; Jobst et al., 2020). The phenomenology of the psychedelic experience is thought to accord with this description of the underlying computational mechanisms. Recall, in a predictive coding scheme, if prediction error can be explained away at lower levels, high-level representations of the model remain stable, as there is no need to update. Under psychedelics, the relaxation of high-level priors means that prediction errors that would usually be explained at lower levels are driven up the predictive hierarchy, resulting in instability in higher-level representations, whereby high-level priors no longer constrain lower-level predictions. At lower doses, this manifests as the phenomenological effects of psychedelics—for example, walls may have the appearance of 'breathing' (Carhart-Harris & Friston, 2019).

This relaxation of high-level priors results in the system adopting a high Bayesian learning rate on sensory evidence (Deane, 2020; Hohwy et al, 2017; Mathys et al, 2014). A low learning rate means there is a greater influence of higher-level priors in determining the resulting posterior, and a high learning rate means there is higher precision on sensory evidence and less constraint imposed by higher-level priors. Appropriately setting the Bayesian learning rate—the precision on sensory evidence—is crucial for the system approximate Bayesian inference over time, as an overreliance on prior expectations leads to a failure to learn from sensory evidence, and an overreliance on sensory evidence can lead the system to "overfit"—essentially, find patterns in noise. The perceptual effects of psychedelics can be characterized as 'rampant' overfitting of sensory evidence (Deane, 2020)—where the system cycles through candidate hypotheses to explain the influx of highly precise prediction error ascending the cortical hierarchy.

Another feature of psychedelic phenomenology is the almost mystical quality that sensory impressions take on. For instance, in a direct comparison between psilocybin and DXM (a non-serotonergic psychedelic) experiences, psilocybin produced “greater visual, mystical-type, insightful, and musical experiences” (Carbonaro, Johnson, Hurwitz, & Griffiths, 2018, p. 1).

Consider Aldous Huxley’s descriptions of the mescaline experience in *The Doors of Perception* (1952):

> “I looked down by chance, and went on passionately staring by choice, at my own crossed legs. Those folds in the trousers— what a labyrinth of endlessly significant complexity! And the texture of the gray flannel—how rich, how deeply, mysteriously sumptuous!” (p. 39)

> “The books, for example, with which my study walls were lined. Like the flowers, they glowed, when I looked at them, with brighter colors, a profounder significance.” (p. 24)

On the current account, this quality of significance which accompanying the perceptual effects of psychedelics can be understood as the system inferring high epistemic value due to the high precision on sensory evidence. This is because: “the better the precision on the prediction error, the higher the learning rate; that is, the more we trust the quality of the evidence the more we should learn from it” (Hohwy, 2017, p. 76). Under the view of affective experiences broadly described earlier as inference about ‘how well am I self-evidencing?’—the positive emotions in psychedelic experiences: “exhilarated elation with unmotivated laughter, deep feelings of peace, exuberant joy, and hedonistic pleasure” (Preller & Vollenweider, 2016, p. 236)—could put down to the greater than expected epistemic value associated with the current policy. This point is relevant for the affective characterisation of ego-dissolution to come.

## 12. Psychedelic-induced ego-dissolution

Deane (2020) characterizes psychedelic induced ego-dissolution as resulting from a failure of sensory attenuation (see Girn et al, 2020; for recent empirical support for this hypothesis). Recall, predictions of the sensory consequences of actions (‘corollary discharges’) allow the system to differentiate between endogenous (self) and exogenous (other) causes of sensation, such that unexpected sensation to be attributed to external causes. Under a view of phenomenal selfhood as allostatic control, the sense of being an agent arises from inferring oneself to be an endogenous cause of sensation; that is, determined by the predictability of action-outcome contingencies. A number of disruptions in self-experience have been accounted for in these terms.  For instance, the symptomatology of schizophrenia —such as thought insertion, where patients report feeling that their thoughts are not their own— have been understood in terms of a failure of these sensory attenuation mechanisms (Ford & Mathalon, 2005; Frith, 2005; Rösler et al., 2015; Thakkar, Mathalon, & Ford, 2021). Here, the system fails to attribute self-generated outcomes to endogenous rather than exogenous causes, manifesting phenomenologically to the agent as a loss of agency over their thoughts (Gallagher, 2004; O'Brien & Opie, 2003; Stephens & Graham, 1994). In the case of voice-hearing this can result in attribution of inner speech to an external source such as another agent (Ford, Gray, Faustman, Roach, & Mathalon, 2007; Ford & Mathalon, 2005).

Corollary discharges—as predictions of the sensory consequences of actions— act to cancel out self-generated sensory outcomes via sensory attenuation. Unexpected consequences are then attributed to exogenous rather than endogenous causes. This means that the more sensory prediction error is generated, the more likely it is that an action or thought has external as opposed internal or endogenous causes (Corlett et al., 2019; Frith, 2005). Deane (2020) notes that under the REBUS model of the action of psychedelics, the influx of both exteroceptive and interoceptive prediction error means that the outcomes of (mental) actions become radically unpredictable. As such, the system ceases to posit itself as an endogenous controller of sensation (and as a causally efficacious agent) as a result—manifesting phenomenologically as ego-dissolution. In the account of thought insertion above, the thought was attributed to 'other' rather than self, due to not being inferred to be self-generated, based on a failure of these mechanisms. Ego-dissolution here is being understood in terms of similar mechanisms to the example of thought insertion described above but is experienced as a more global dissolution of selfhood due to the influx of unpredictable inputs from across the cortex, as opposed to being isolated to certain activity.

### Affective tone

While ego-dissolution is described as being devoid of self-consciousness, it is nonetheless described as a highly conscious state, characterized by affective extremes. Carhart-Harris & Friston (2020) distinguish between "complete ego-dissolution" —a state of "complete surrender, associated bliss, and union with all things" (Carhart-Harris & Friston, 2019, p. 321); and "incomplete" ego-dissolution—a state characterized by intense fear, anxiety, and distress.

'Complete' ego-dissolution can be understood on the current account to be underpinned by two closely related computational mechanisms. The first relates to pragmatic value and prior preferences. Recall, on the account of self-modelling proposed here the inference on allostatic control tunes precision on expected free energy. For instance, this could be higher precision on a particular prior preference (consider the example of pain perception given earlier). Precision on unfulfilled prior preferences can be a persistent source of suffering to the system—one such example being chronic pain, where chronic pain is underpinned by high precision on a prior preference that is unable to be fulfilled through action (Hechler, Endres, & Thorwart, 2016). On the view that action arises from minimising the discrepancy between the actual (inferred) current state and the desired state, relaxation of the constraining influence of high-level priors means they cease to structure consciousness to engage the organism in their fulfilment, and as such, end their associated suffering. High-level priors constraining more domain general affective states such as mood (Clark, Watson, & Friston, 2018) would also be relaxed under the REBUS model. This connects closely to the therapeutic potential of the experience: "psychedelics work to relax the precision weighting of pathologically overweighted priors underpinning various expressions of mental illness" (Carhart-Harris & Friston, 2020, p. 1). Deane (2020) highlights that the lessened influence of prior preferences accords with descriptions of the phenomenology of ego-dissolution, for instance: "It felt as if 'I' did no longer exist. There was purely my sensory perception of my environment, but sensory input was not translated into needs, feelings, or acting by 'me'" (unpublished online survey data quoted in Millière et al., 2018, p. 7).

There is another reason 'complete' ego-dissolution may be characteristically ecstatic. Inference about allostatic control is not just about realising the pragmatic affordances of action, but also in maximizing the epistemic value associated with a given policy. Recall, in normal functioning, precision on sensory information would track the expected epistemic value of sensory inputs. In the psychedelic state, the relaxation of high-level priors, and corresponding increase in sensory precision, means the system infers infer that the current state is realizing great epistemic value (See 'Psychedelics and Insight' in Carhart-Harris & Friston, 2020).

## 13. Responding to the selflessness challenge

We can now return to the question of whether psychedelic induced ego-dissolution threatens active inference theories of consciousness grounded in self-modelling. To recap, Millière (2020) argues that some states of consciousness are "totally selfless, insofar as they do not involve any form of self-consciousness." (Millière, 2020, p1) As such, these states refute versions of the necessity claim: the view that self-consciousness is necessary for consciousness in general. In light of the preceding discussion of ego-dissolution in the predictive brain, I will address this challenge in this section. However, I do not intend to argue whether or not ego-dissolution should be understood as "totally selfless" or simply "partially selfless" states. Instead, I will argue under the account of ego-dissolution proposed in Deane (2020) and in this chapter, these selfless experiences do not undermine active inference theories of consciousness in terms of deep self-modelling. I will do this by showing that whether these states are cast as "totally selfless" or "partially selfless" is inconsequential for an active inference theory of consciousness, as on either reading the mechanisms underlying consciousness itself are explicit.

First, let's consider the view of ego-dissolution as a 'partially selfless' state. A view of ego-dissolution as a partially selfless state of consciousness presents no problem to the subjectivity theorist, and, as such, no problem to an active inference approach to consciousness in terms of phenomenal self-modelling, because it could be that whatever aspect of self-consciousness is missing, it is not the relevant aspect for consciousness. In light of the current discussion, the grounds for this view are that the affective inference that remains intact in the state of ego-dissolution is understood as a kind of self-consciousness. On the current account, this view might appeal to the fact that an inference about agentive control and affective inference can come apart: while the system may cease to posit itself as an endogenous controller of sensation (and as such cease the phenomenology of being an autonomous agent), a domain general inference about allostatic control (how well am I self-evidencing?) remains intact. For instance, in 'complete' ego-dissolution the combination of relaxation of high-level prior preferences and the high epistemic value, means the system infers itself to be in a state of high allostatic control—the inference of subjective fitness understood to underpin the affective dimension of phenomenal selfhood remains present. If this affective state is taken to be a kind of self-consciousness, then even in this radically stripped back experience, the system is still in a state of self-conscious and as such this state is compatible with the necessity claim.

It is reasonable to assume that many will not be satisfied with this characterisation self-consciousness. For instance, on the definition of self-consciousness as consciousness of oneself *as oneself* (Millière, 2020; Smith, 2017), it seems reasonable to conclude that psychedelic-induced ego-dissolution is best understood as being totally selfless, as argued by Millière (2020) and Letheby (2020). It may be the case that whether ego-dissolution is understood as a state totally devoid of self-consciousness boils down to how self-consciousness is defined. Millière (2017; 2020) notes that the disagreements about the necessity claim and the typicality claim may hinge on terminological variation (Guillot, 2017), due to the polysemy of "self-consciousness" and "sense of self". Crucially, settling this debate is inconsequential for whether ego-dissolution could be a problematic case for an active inference theory of consciousness in terms of self-modelling, because even in the case that we accept these cases as instances of totally selfless experience, we can make a distinction between computational self-modelling and phenomenal self-modelling.

Limanowski & Friston (2020) argue that selfless states can be understood as "(rare) cases in which normally congruent processes of computational and phenomenal self-modelling diverge" (p12). On the account of ego-dissolution put forward in this chapter, the phenomenology of ego-dissolution (complete or incomplete) is underpinned by a very particular inference about allostatic control. As such, the affective inference argued to underpin consciousness remains intact in the state of ego-dissolution. If self-consciousness is truly absent in states of drug-induced ego-dissolution—that is, if the affective inference present in these states is deemed not to qualify as self-consciousness due to not being understood as a representation of oneself *as oneself*— then the active inference approach to consciousness put forward in this chapter simply doesn't qualify as a subjectivity theory, and so is not vulnerable to the selflessness challenge.

The present account puts subjective valuation as the most basic constitutive feature of a conscious experience—sensation is always infused with what it *means* for the organism: "not everything that happens to us enters our awareness, not by far; but everything that does is not merely registered but also felt." (Kolodny, Moyal, & Edelman, 2021, p. 4). This fact is brought into sharp relief in the account of psychedelic-induced ego-dissolution, as affective experience remains even when all the other structuring features of experience are extinguished. The upshot of this view is that affective valence can be understood as the most fundamental part of conscious experience (Damasio, 2019;

Damasio & Carvalho, 2013; Damasio, 1999; Man & Damasio, 2019; Panksepp, 1998, 2005, 2008) . This is a view that dates back at least as far as George John Romanes:

> "The raison d'être of Consciousness may have been that of supplying the condition to the feeling of Pleasure and Pain." (Romanes, 1888, p. 111)

Grounding consciousness in basic affectivity also has a precedent in the literature on the free energy principle and consciousness, which is particularly consonant with the current picture. Mark Solms has argued, partially based on evidence of consciousness in decorticated animals and congenitally decorticate (hydranencephalic) humans, that:

> "Consciousness itself is affective. Everything else (from motivation and attention, leading to action and perception, and thereby to learning)—all of it—is a functional of affect. Affect obliges the organism to engage with the outside world." (Solms, 2019, p. 12)

Moreover, Solm's view seems aligned with the view of ego-dissolution as only partially selfless states, and with the view proposed here that higher-layers of the phenomenal self-model structure consciousness:

> "Affect just is a self-state (and through feeling—i.e., precision optimisation—it necessarily generates consciousness itself), which activates (selects) salient perceptual representations, which eventually include cognitive re-representations of the self." (Solms & Friston, 2018, p. 17)

On this view, feelings in the form of subjectively felt valence are the most basic constitutive phenomenal states, they pervade all of experience, guiding the organism to fitness promoting states (Inzlicht, Bartholow, & Hirsh, 2015; Kolodny et al., 2021). While in typical experience, subjective valuation functions to fine-tune learning (Eldar, Rutledge, Dolan, & Niv, 2016) and regulate behaviour we can see the same mechanisms in place in atypical experience, such as the psychedelic state.

## 14. Conclusion

This chapter has argued that phenomenal consciousness is best understood within predictive processing in terms of the deep self-models inherent in the active inference framework. On this account, subjectivity is structured by a 'deep control model'—a hierarchically deep self-model that tracking the temporally deep endogenous control of self-evidencing outcomes. Higher-levels provide deep contextualisation (interoceptive inference) of afferent signals from the body (Miller & Clark, 2017), tuning the organism to adaptive opportunities for action. Two objections to this view have been considered: i) that the core characteristics of consciousness in predictive processing is underspecified, and as such cannot inform which systems are conscious, and; ii) the challenge of psychedelic-induced ego-dissolution. I have argued neither of these objections is troubling for an active inference theory of consciousness, and as such active inference is a very promising framework for consciousness science.

# Chapter 6: Expecting Action: Predictive Processing and the Construction of Conscious Experience

Predictive processing has begun to offer new insights into the nature of conscious experience – but the link is not straightforward. A wide variety of systems may be described as predictive machines, raising the question: what differentiates those for which it makes sense to talk about conscious experience? One possible answer lies in the involvement of a higher-order form of prediction error, termed expected free energy. In this chapter we explore under what conditions the minimization of this new quantity might underpin conscious experience. Our speculative suggestion is that Expected Free Energy is relevant only insofar as it delivers what Ward, Roberts & Clark (2011) have previously described as a *sense of our own poise over an action space*. Perceptual experience, we will argue, is nothing other than the process that puts current actions in contact with goals and intentions, enabling some creatures to know the space of options that their current situation makes available. This proposal fits with recent work suggesting a deep link between conscious contents and contents computed at an 'intermediate' level of processing, apt for controlling action.

## 1. Introduction

Predictive processing offers new insights into the nature, possibility, and structure of conscious experience. By seeing experience as a construct that merges prediction and sensory evidence, we begin to see how minds like ours infer the structured world. This is a world built around two core necessities – the need to select apt world-engaging actions (such as reaching fora glass of water) and the need to maintain the inner milieu within the bounds of human viability. The two are clearly linked, though the space of apt actions soon outruns the space of actions whose purposes are directly related to keeping us within our species-specific window of biological viability.

Embodied agents are able to explore the space of possible actions by minimizing expected future prediction error (expected free energy, or EFE). Recent speculations concerning possible links between predictive processing and conscious human experience make essential reference to this quantity, EFE, which is different from standard variational free energy (Friston et al. 2016; Parr & Friston 2019; Millidge et al. 2021). We ask under what conditions the minimization of EFE underpins the emergence of conscious experience. Our suggestion is that minimizing EFE is not sufficient for the presence of conscious experience. Instead, EFE is relevant only insofar as it delivers what one of us (see Ward, Roberts & Clark, 2011) has previously described as a sense of our own poise over an action space.

In this chapter, we reconstruct the notion of 'knowing poise over an action space' as implying an agent-inspectable policy space in which candidate policies afford differing actions. Agent-inspectability is unpacked as fine control over the precision-weighting system enabling an agent to launch and assess multiple simulations of possible futures, so as to optimize contact with their own preferences and intentions.

Perceptual experience, we conclude, is nothing other than the process that puts current actions in contact with goals and intentions, enabling some creatures to know the space of options that their current situation makes available. It is the shape of the controllable action-space that determines the 'grain' of conscious experience – why we perceive cups, shapes, and colours but not lower-level best-guesses such as the location of 'zero-crossings'. Experience is populated by 'intermediate level representations' (Marchi & Hohwy (2020)) that capture actionable properties and features, and what is actionable reflects contingent facts about the creature, it's learning history, and its environmental niche.

In section one we give an overview of how active inference can be understood in terms of a formalisation of allostasis, with a focus on how the sense of agency and phenomenal self-modelling is implicit within the active inference framework. Section two we consider what such a broadly applicable story can tell us about the much rarer phenomenon of subjective experience, and consider the limited cases in which there is the development of a hierarchically deep self model and the ability to minimize expected free energy, as a potential marker for the emergence of consciousness. In section three we draw upon the action space account, developed by Ward, Roberts and Clark (2011), connecting it (section 4) to recent work on predictive processing and the spatiotemporal resolution of human action (Marchi & Hohwy 2020). In section five we then deepen this connection by showing that free energy minimization grounds the action space story in basic homeostatic imperatives and architectures of control that support a wider repertoire of subsidiary goals and action policies. In section six we pull all these strands together under the banner of 'generative entanglements' (Clark 2019) as the mechanistic basis of conscious experience. It is these entanglements, we then argue (section 7) that are responsible for our own Cartesian intuitions – the very intuitions that make many doubt the possibility of a satisfying mechanistic account of consciousness.

## 2. Allostasis and Expected Free Energy

Central to the free-energy principle is the premise that organisms survive by maintaining their internal states within viable homeostatic bounds (Cannon, 1929; Friston, 2012a). Interoception - 'sensing the body from within' (Craig, 2002) - is essential for the maintenance of viable internal states. Increasingly, predictive processing mechanisms thought to underlie perceptual inference in exteroception are being applied to understand how the brain performs interoceptive inference (Barrett & Simmons 2015; Seth 2013; Pezzulo 2014). Predictive models on the state of the body allow the system to act to bring internal states—'essential variables'—into reasonable bounds. For instance, when the brain infers the state of the body to be outside the bounds of a healthy body temperature, it can engage corrective autonomic reflexes—such as perspiring when temperature is too high.

Creatures that track regularities across longer timescales gain the advantage of being able to anticipate dyshomeostatic conditions before they arrive, and as such are able to act so as to avoid these outcomes. This anticipatory action or 'predictive regulation' is known as allostasis (Corcoran & Hohwy, 2017; Pezzulo, Rigoli, & Friston, 2015; Schulkin & Sterling, 2019; Sterling, 2012). Allostasis allows the system to go beyond stimulus driven and reflex based corrective mechanisms. Central to the concept of allostasis is that organisms maintain "stability through change". Allostasis moves beyond the concept of closed-loop control of homeostatic set points by introducing flexible parameters that can change according to context, such as the anticipatory physiological response to a threatening situation, giving an organism the energetic resources to anticipate and avoid dyshomeostatic outcomes.

Pezzulo et al (2015) posit that allostatic action selection is underpinned by hierarchical generative models that apply prior beliefs to map actions or sequences of actions to interoceptive outcomes over time. This mapping of policy-dependent outcomes allows an agent to infer the actions that bring about favourable interoceptive, proprioceptive or exteroceptive outcomes - such as realising a state of satiation through a certain sequence of actions. Higher hierarchical levels subtending more temporally deep or extended policy dependent outcomes act to contextualise and guide lower levels. While it is thought that most fundamentally this is related to realising self-evidencing interoceptive states over time, the same machinery can be applied such that an organism can not only act on affordances, but can act to bring about future affordances (Pezzulo & Cisek, 2019).

Active inference can then be understood as a formal articulation of allostasis whereby control problems are cast in terms of a model-based inference about the best action plans (or policies). This is known as *planning as inference* (Botvinick & Toussaint 2012, Kaplan & Friston 2018) under the free energy principle (Friston, 2019a). In this scheme, the planning and execution of actions become a problem of inference, where candidate actions are scored with respect to their expected free energy—the average variational free energy the organism expects to accrue in pursuing a given policy. The policy with the least expected free energy is that which is selected. Importantly, the expected free energy can be decomposed into the pragmatic and epistemic affordances of action—which means the agent can balance fulfilling prior preferences alongside the epistemic goal of reducing uncertainty about the causal structure of the world (Fristion 2016).

Self-modelling mechanisms permeate active inference. For instance, action initiation involves systematic misrepresentation of the state of the self: in moving, the organism predicts itself to be in the sensory state which corresponds to the completed movement (Wiese 2017). This involves *disattending* (Limanowski, 2017), that is, lowering precision—about the current sensory evidence about the position of the arm, and attending (allocating high precision) to the desired state (my arm has moved). In other words, all that is required for intentional movement is a specification of the desired sensorimotor endpoint of a movement and motor reflexes bring the motor plant to that equilibrium or setpoint (Adams et al., 2013; Brown et al., 2013; Kilner, Friston, & Frith, 2007; Parr, Rees, & Friston, 2018).

The system also attenuates sensory evidence of the sensory consequences of self-generated movements. In classic motor control, this was understood in terms of a 'comparator model', whereby the system compares expected and actual consequences of action. Monitoring the mismatch between actual and potential consequences of actions was originally argued to underpin refinement of motor commands, and has since been invoked to understand the origin of the sense of agency (Hohwy & Michael, 2017; Lukitsch, 2020). In active inference, this is simply part of the top-down prediction about the sensory consequences of actions, but the principle remains that in monitoring the ongoing (mis)match between actions and outcomes, the system can infer its endogenous control over sensation via action.

In selecting actions and action policies with the least expected free energy, an organism has to infer *itself* as able to bring about the consequences of the action — as an endogenous controller of sensation. While perception in active inference can be understood in terms of inference about hidden states of the causes of sensory signals, action selection requires inference about transitions between states contingent on actions. Importantly, this inference about how sensation is controlled sensation via action tracks both proximal action consequences; for example the sensorimotor contingencies involved in turning over a tomato, and  distal ones; the abstract future consequences associated with a given action or action policy, where the degree of abstraction increases with temporal depth (Pezzulo, Rigoli, & Friston, 2018). This inference about control, inherent in active inference, has been associated with the phenomenology of being a *self* or an agent in the world (Hohwy & Michael, 2017; Limanowski & Friston, 2018; 2020), and has been dubbed 'agentive control' (Deane, 2020; Deane et al, 2020)

## 3. Conscious systems

From steam engines to stock traders (automated or otherwise) regulatory systems are everywhere. Anticipatory regulatory systems may be less common, but we can plausibly count colonies of Burkholderia  proteobacteria (Goo et al., 2012), jumping spiders (Schomaker, 2004), and self-driving cars among their number. One of the central offerings of the active inference formalization of allostasis – and of cybernetic formalizations of homeostasis that preceded it (Ashby, 1952) – is the means of abstracting beyond the organism, to identify these same dynamics across a wide range of physical systems. In this spirit, the Free Energy Principle has been applied not only to brains and bacteria, but also to coupled pendulums (Kirchoff et al., 2018), the Watt governor (Baltieri, Buckley & Bruineberg, 2018), and an oil drop suspended in water (Friston, 2019b) both in support, and as criticism, of Friston's assertion that it stands as a theory of "every 'thing' that can be distinguished from other things in a statistical sense" (Friston, 2019a).

So what does active inference have to say about the distribution of a conscious experience across this diverse array of systems and scales? To take the dynamics of active inference as being the hallmark of a conscious being would result in an extremely liberal view of what suffices for the attribution of sentience indeed. Conservatism about consciousness aside, we know from our own lack of experience that the majority of our regulatory processes, from shivering when cold to going through a morning routine on autopilot, need not involve conscious awareness.

One possibility is to argue that active inference and predictive processing in themselves are simply not about consciousness at all. In taking this route Anil Seth and Jakob Hohwy (2020) propose that it is precisely because predictive processing is not itself a theory of consciousness, that it provides an ideal foundation for building such a theory. The emerging consensus in active inference is that consciousness is grounded in the self-modelling processes that are inherent in the action selection and precision control mechanisms in active inference. In this way, consciousness emerges only in systems that minimize expected free energy by the selection of action policies over time. However, this may still provide at most a necessary condition on conscious experience. As Friston (2018) points out, not all active inferrrers are equal. Within the general class of free energy minimizing systems we can identify those with a more temporally-thick model, granting the capability to infer far into the future; with the counterfactual-depth needed to mentally explore the consequences of possible non-actual actions, and the development of a self-model granting the capability to involve one's own projected future needs into that calculation. Such capacities, he proposes, are just what is needed to sort those that are conscious, from those that are not:

> "One could then describe systems that have evolved thick generative models (with deep temporal structure) as agents. It now seems more plausible to label these sorts of systems (agents) as conscious, because they have beliefs about what it is like to act; i.e., just be an agent. Furthermore, because active inference is necessarily system-centric the self-evidencing of motile creatures can only be elevated to self-consciousness if, and only if, they model the

> consequences of their actions. Put simply, this suggests that viruses are not conscious; even if they respond adaptively from the point of view of a selective process. Vegans, on the other hand, with deep (temporally thick) generative models are self-evidencing in a prospective and purposeful way, where agency and self become an inherent part of action selection" (ibid).

Consciousness, then, emerges in systems that evaluate deep and counterfactual rich models of the world. These requirements of temporal depth and counterfactual thickness of a generative self-model do seem to track the behavioural capacities that mark the transition to increasingly sophisticated forms of life – the difference between an E.Coli bacterium's ability counteract a drop in glucose levels by breaking down glycogen, versus my ability to prepare a sandwich in anticipation of becoming hungry later this afternoon. They do not, however, provide a clearcut answer to whether something becomes conscious or not – only a graduated scale against which a particular system might be assessed. As Friston, Wiese & Hobson (2020) acknowledge the view of consciousness as emerging from increasing hierarchical modelling layers, "entails that there is only a gradual difference between some non-conscious and conscious systems, and that consciousness is a vague concept."

But suppose we are willing to accept this – willing to consider the question "conscious or not" as one that admits of degree. Having a temporally-thick, counterfactually deep, self-model seems an obvious prerequisite for the presence of higher-order self-consciousness, but it is not immediately clear why it would be necessary for the emergence of more basic phenomenal properties—such as the immediate visual experience of an apple on the table. In order to explain why we believe that it is then, we need to step away from the FEP, and PP for a moment, to take another look at the content of such simple visual experiences.

## 4. The Action-Space Account of Perceptual Content

We know from our own case that even in creatures like ourselves, visual experience is typically not required for the co-ordination of a surprisingly sophisticated repertoire of behaviours. There is a wealth of empirical evidence that visual awareness is unnecessary to visually-guide behaviors, such as pointing,  tracking and reaching – all of which ordinary participants are able to perform without conscious perception of their target (Bridgeman, Kirsch & Sperling, 1981; Goodale, Pélisson & Prablanc, 1986; Castiello, Paulignan, and Jeannerod, 1991).

This becomes particularly striking in neurological disorders such as visual agnosia  and action-blindsight (Danckert & Rossetti, 2005) where damage to areas of the brain involved in visual processing significantly disrupts phenomenology without equivalent impairments to visual action-guidance. For instance, the well-studied visual-agnosic,DF (Whitwell, Milner & Goodale, 2014; Milner and Goodale, 2006) who  lacks visual awareness of  the size or shape of a slot in front of her eyes, yet, when instructed to do so, can post a letter through that same slot with perfect ease.[13]

So conscious visual perception may be significantly impacted without impairing online visual guidance of pre-established regulatory-routines – unconscious vision seems to handle this well enough on its own. Instead Ward, Roberts & Clark (2011), propose the 'action space' account of perceptual experience, which argues that:

> *"... what counts for (what both explains and suffices for) visual perceptual experience is an agent's direct unmediated knowledge concerning the ways in which she is currently poised (or, more accurately, the way she implicitly takes herself to be poised) over an 'action space." (Ward, Roberts & Clark, 2011 p.383)*

The absence of this direct awareness of the range of action-routines currently available to her is manifest in DF's behavioural capacities. To talk of her impairment as merely perceptual, as Milner and Goodale (2006) initially did when using DF's case to support the division of visual processing into unconscious 'vision for action' and conscious 'vision for perception,' misleadingly implies that the capacity for action is entirely unaffected by the loss of conscious perception. Such a strong division overlooks not only the fact that DF's inability to produce utterances appropriate to incoming visual stimulation is itself a kind of action, but also that she is further unable to: indicate the width of the slot with her hands; match the orientation of a letter to that of the slot without posting it; take the initiative to post the letter without prompting; or to scale her grasp to a briefly presented object after any delay.

[13] Our argument does not rest on the claim that D.F lacks visual experience altogether – which is contentious. The point is only that her visual experience changed radically upon the damage to her ventral stream and that such a difference in experience is not 'purely perceptual', but is rather inseparable from consequent impairments to action-planning capacities.

DF can use visual input to guide the ongoing unfolding of a pre-specified action towards a target, once she is involved in this action. She is unable, however, to plan potential actions, to anticipatorily move her body to prepare for the execution of an action, or to perform an action in relation to visual information that is no longer directly available. What she lacks, Ward, Roberts, and Clark argue, is the ability to automatically integrate this visual input with her background knowledge and goals to gain knowledge of the way in which she is poised over an action space.

Due to this integration between ongoing sensitivity to changing environmental affordances, and one's prior body of understandings and intentions, this 'immediate knowledge of one's poise over an action space' is, Ward, Roberts and Clark argue, more sophisticated than the kind of practical knowhow deployed in the execution of a pre-established motor routine. Yet they also stress that it is not to be understood in intellectualized terms, as involving fully-fledged, detachable and context-neutral, conceptual abilities either. To require this would seem to place the bar for visual experience unacceptably high. Besides, it is perfectly possible to reason through the actions available to one at the conceptual level, such as what show to watch on Netflix this evening, with no corresponding visual phenomenology.

PP/Active Inference we argue, provides the means to operationalize this form of online action planning – as distinct from both conceptual reasoning and online action control – by framing our cognitive operations in terms of the attempt to minimizing error across multiple timescales, via a hierarchical predictive architecture. Mere online control amounts to the sending down of a fixed prediction that guides action to bring one's incoming sensory signal in-line. Detached, non-visual reasoning corresponds to revising higher levels of the model to reduce internal incongruities – absent the transmission of any action-eliciting predictions, and consequent feedback from, the sensorimotor periphery. In contrast, when an "agent's perceptual sensitivity is such as to automatically mesh with her capacities for intentional activity" (Ward et al. 2011) we have a circular feedback loop spanning higher and lower levels, with longer term predictions constraining the operations of the lower levels, while continuous/high-precision error signals at the lower levels seep up to trigger adjustments at these higher levels in turn.

## 5. Intermediate Level Processing

Within these circular feedback loops, predictions that depict objects and properties at a certain grain or level seem to play a special role. Marchi and Hohwy (2020) address what they call the 'scope question' viz 'what we are conscious of, given that we are conscious at all' . They make a compelling case for the claim that 'intermediate level representations' play a special role in determining the scope of conscious contents within an active inference framework. These representations are constructed within the predictive hierarchy at a level that is neither too abstract nor too fine-grained to guide the selection of policies for basic action. According to their picture, it is the spatiotemporal resolution of typical human actions that determines this level, which might thus vary for different organisms. The neural realizers of conscious contents, they argue, is determined by the role they play in selecting actionable policies.

Their proposal builds on previous ideas concerning the privileged status of certain intermediate-level representations, beginning with Jackendoff (1987) and continuing though Prinz (2000), Koch (2004), and Prinz (2012, 2017). The general idea is that intermediate level representations sit between characterizations that are too abstract to determine one action rather than another, and those that are too low-level. They are thus defined relative to the basic action dispositions of the organism – dispositions that, in humans, might include "grasping, kicking, and turning the head" (Marchi & Hohwy, 2020, p.8).

When we encounter a world of actionable objects and states of affairs, we do not experience every low-level nuance in our own processing. For example, we do not experience the computations of 'zero-crossings' that seem to underlie edge and boundary detection, or the multiple low-level hypotheses that must be varied and updated as we look at an object from various angles . Instead, all we see is a bound, shaped object, rotating in egocentric space, Nor do we visually experience highly abstract properties such as pure object-hood.

Strikingly, the higher and lower bounds of the experiential realm seem to reflect the kinds of information suited to the selection of one kind of basic action over another. Marchi and Hohwy offer a characterization of basic action in terms of a repertoire of organism-specific possibilities such as reaching out with the hands, turning the torso, and so on. Their idea is that some levels of the generative model preferentially depict a possibility space for organismically basic action. It is only at those privileged levels that precise actionable polices can be inferred.

This proposal is a good fit, we suggest, with the action space account. To succeed in our intentional plans and projects, we need to infer and implement actionable policies. The phenomenal realm, if their proposal is on track, is constructed so as to enable us to encounter our world in a way that is ready-parsed for the kinds of basic action we can perform. Precise policies (in the active inference sense of 'precise') are ones that will deliver high amounts of reliable prediction error, so that minimizing those errors controls fluent successful action. Neither the very low-level nor the very highest-level control states (the ones that determine tiny bodily nuances or drive long-term projects and goals such as writing a book) fit this bill. Longer-term policies such as writing a book are precise and actionable only to the extent that they are composed of, or reliably give rise to, sequences of policies that engage basic actions in this way.

Optimal performance demands that the selection of local action policies is consistent with the availability of precise control. It is this demand that explains the privileged status of the intermediate level information that seems to populate phenomenal experience. The scope of the phenomenal realm, they argue, is delimited by the level of the generative model at which the space of actionable policies can be safely explored. Flagging that level stops us from constantly inferring policies that we cannot implement.

As Marchi and Hohwy put it:

> "a conscious agent is conscious of the hypotheses that are flagged at the appropriate resolution for optimal inference of policies allowing efficient and successful performance of sequences of basic actions (control states)." (p.17)

In the case of DF, we suggest that this flagging or highlighting has in some way broken down. Although she can, if prompted, engage the letterbox with a posting action, her visual encounters (thanks to the damage to her ventral stream) do not present her with a rich realm of flagged possibilities for action. This makes sense if the ventral stream damage is depriving her of many of the kinds of precise information she would normally expect. It would be an interesting exercise to attempt a full reconstruction of this using the resources we have now assembled, but this is a project we must leave for another day

## 6. Giving the Action Space an Allostatic Foundation

Like Marchi and Hohwy's picture, and older cousins such as sensorimotor contingency theory (Noe & O'Regan, 2003) the action space view ties our visual experience to our capacities for bodily action. Unlike on the former account, however, the content of our action space is constructed at a level more coarse-grained than exact sensorimotor trajectories, and more selectively organized around our particular goals and interests. As Ward, Roberts & Clark. write: "An action space, in this specific sense, is to be understood not as a fine-grained matrix of possibilities for bodily movement, but as a matrix of possibilities for pursuing and accomplishing one's intentional actions, goals and projects" (Ward et al, p 383). In moving to explain perceptual content in terms of the content of a space of intended actions, it lacked, however, an account of how these goals and the structure of the action spaces develops, of the principles guiding this selectivity and coarse-graining. As such it appears to fall foul of what Susan Hurley (1998) termed, 'the Myth of the Giving' – that is, of attempting to explain perceptual content in terms of 'just more content' as though the content of our intentions could be taken as explanatorily basic. Marchi and Hohwy's treatment of the scope question marks real progress here. In the remainder of this chapter, we seek to go further still, to sketch a view of the fundamental nature of the conscious mind. Our aim here is to enrich the action space/intermediate level processing picture in ways that highlight the role of affect and layered control,  displaying patterns of both continuity and discontinuity with the basic allostatic profile of life.

By treating neither perception nor action as more primitive than the other, but instead, in the spirit of Hurley's (2001) proposal, understanding both as interdependent means of control, cybernetic/sensorimotor PP provides a basis for the construction of higher-level action spaces, as founded upon on the basic and fundamental imperative to keep our allostatic self within viable bounds. Beginning with the phylogenetically wired-in prediction that my blood sugar level should be 70 to 100 mg/dL, I can also learn the regularity of this dropping (prediction error increasing) as 1pm approaches.  Through the trial and failure of ontogeny I can learn that not only does releasing glucose stores reduce this prediction error, but also that eating a meal at 12pm can prevent this prediction error from arising at all. Thus I begin to predict daily meal-eating at 12pm. If I've not started eating as midday approaches, then prediction-error now results. The content of my perception now becomes shaped by  action possibilities which I have learnt will reduce this prediction error – the trajectory towards the opening of the fridge, the tub of soup and the microwave.

We also learn, at what level of detail this action trajectory needs to be enforced in order to result in successful control. A wide array of specific sensorimotor trajectories, corresponding to opening the fridge, will interchangeably minimize prediction error relative to the expectation of the cup being in our hands. Thus, when it comes to exploring potential control strategies, we learn that we do not need to individually explore each and every possible combination of sensorimotor signals. Instead, these details can be condensed into a single course-grained action possibility, that can then be evaluated for its potential to bring our sugar level back within expected bounds. Which trajectories stand out to us for exploration, and at which level of detail, will shift not only with changes in the external world, but also as our internal situation changes. When I'm anticipating a drop in blood sugar levels, it is the pathways to the fridge, the soup, and the microwave that dominate. When I'm tired, my exploration of potential actions skews towards the sofa, the remote control and the television.

This gives us a picture of how the action space account can be augmented by this conception of allostatic control. The agentive control in the allostatic control model is not an impassive observation of the control of sensory inputs via action. Rather, it is concerned with bringing the sensorium in line with the "attracting set"—i.e. self-evidencing outcomes. To do this, as we have noted, involves selecting policies associated with the least expected free energy, where expected free energy can be decomposed into the pragmatic and epistemic value associated with the given policy. This means, in selecting a policy an organism seeks both to fulfil prior preference (for example, become satiated), and also to resolve uncertainty about the world. This bridges the current proposal to accounts grounding selfhood in the body (Allen & Friston, 2016; Apps & Tsakiris, 2014; Limanowski & Blankenburg, 2013; Seth, 2014)

By tracking at how well it is faring in bringing about self-evidencing outcomes, the organism can infer the precision on its own action model—that is, an inference about control in a given context. This precision estimate on the action model—as inference about confidence in control of, for instance, prior preferences, is thought to underpin *affective inference*—why it *feels like* something to be an organism. Affective inference tracks context specific precision on (for instance) prior preferences. For example, violation of the "healthy body condition" expectation (Ongaro & Kaptchuk, 2019) manifests to the system as pain, and the system must act to bring sensations into line with the expectation (where the expectation here is physiological integrity). Importantly, organisms are able to tune the affective tone of the painful percept according to context—as evinced by Bayesian and predictive coding mechanisms pointing to the fact that pain seems to be inferential (Anchisi & Zanon, 2015). Bayesian models of pain perception (Büchel, Geuter, Sprenger, & Eippert, 2014). Painful percepts, on this view, integrate prior beliefs with the current sensory evidence, weighted by their expected precision in the context (Morton, El- Deredy, Watson, & Jones, 2010). Tuning pain perception allows for appropriate motivational salience in the context. Stress-induced analgesia is one illustration of this—the pain of a twisted ankle should not be motivationally salient when trying to outrun a bear.

High-level priors in the generative model can track temporally deep outcomes, and as such, failure to meet an expected *rate* of prediction error reduction over time manifests to the system as negative affective (Joffily & Coricelli, 2013; Kiverstein et al., 2017; Van de Cruys, 2017). A better than expected rate of prediction error over time manifests to the system as positive affect (consider unexpected rewards, either those that fulfil prior preferences like ice cream, or unexpected epistemic rewards—an "aha!" moment for instance).

The upshot of this is what salient to an organism is not constrained to, for instance, the smell of food when hungry. Instead, organisms with hierarchically deep contextualization of interoceptive signals are tuned to appropriate action and engagement with environmental affordances (Pezzulo & Cisek, 2019), and assign appropriate weight to priors and ascending prediction errors across the cortical hierarchy, including context dependent gain control in sensory cortices. Even the smell of hot food may not be salient to an organism engaged in, for instance, finishing a paper ahead of a deadline.

In this way, precision on policies inference about *intentional selection* ('what am I doing')— determines *attentional selection* (What am I seeing?). Attentional orientation – where precision is assigned in the cortical hierarchy, depends on hierarchically deep interoceptive inference on the situation. The control context is similarly vital—inference on what "I can" (Bruineberg, 2017) do determines where attention should shift to update the model to perform allostatic action and behavioural control, according to goals. Inferences about action-outcome contingencies (control), across multiple timescales, informs how precision (attention) is assigned accordingly relative to the organisms inferred control.

This proposal thus delivers on both the sensorimotor and the affective dimensions of consciousness. As Ullman et al. [2017, 649] put it, 'We implicitly but continually reason about the stability, strength, friction, and weight of the objects around us, to predict how things might move, sag, push, and tumble as we act on them.' (Ullman, Spelke, Battaglia & Tenenbaum 2017) Experiencing the world as made made up of objects and states of affairs is accounted for by the fact we have counterfactually rich and temporally deep expectations of the unfolding sensory flow—control of sensation contingent on actions—were we to interact with the world in certain ways, bridging the predictive processing approach to sensorimotor theories of consciousness (Seth, 2014; Noe & O'Regan, 2003). At the same time, integrating these expectations of control with inference about (temporally deep) expectations of self-evidencing outcomes, furnishes this proposal with the affective dimensions of consciousness, such that we encounter "a structured world apt for action and intervention, and inflected at every level, by an interoceptively-mediated sense of mattering, reflecting 'how things are for me as an embodied agent"' (Clark, 2019, p7).

## 7. Generative Entanglements

Pulling these strands together yields a concrete proposal concerning the machinery responsible for conscious experience. Deeply entangled with our grip on the outside world, an inward-looking (interoceptive) cycle targets our own changing physiological states – states involving the gut, viscera, blood-pressure, heart-rate, and the whole inner economy underlying hunger, thirst, and other bodily needs and appetites. As our bodily state alters, the salience of various worldly opportunities (to eat, for example) alters too. That means I will also act differently, harvesting different streams of information. Philosophers and psychologists talk here of 'affordances' (Bruineberg & Rietveld, 2014), where these are the opportunities for action that arise when a certain type of creature encounters a certain kind of situation – for example, a hungry green sea turtle encountering a nice patch of algae discovers an affordance for eating. But the sea turtle that has just eaten may not find the next patch of algae quite so attractive. Such creatures orient towards the changing value of different affordances given their changing bodily needs.

That already captures a form of very basic sentience. We can think of basically sentient beings as those whose neural model of the (organism-salient) world is in constant two-way communication with their own changing physiological state. Such creatures will perceptually encounter a world fit for action, in which what actions are selected depends heavily upon their current and on-going bodily state and needs. But this falls short, we have argued, of delineating the conditions responsible for true conscious awareness.

What's missing is something just a little bit 'higher order'. The creature we just imagined is in touch with its world, in a way that brings together bodily (allostatic) needs and the opportunities for action made available by the sensed environment. But to truly experience that world, the information available to drive action needs to be in some elusive sense 'available to the creature in question'. We have tried to unpack this notion by suggesting that the creature needs not simply to act, but should find itself confronted with an action space – a perceptual array that affords multiple responses, and that (in so doing) is in touch with capacities for planning and intentional action.

This inserts a kind of gap between sensory stimulation and action, one that sometimes results in characteristic behaviors such as pausing to ponder what to do next. It may be that the sea-turtle finds itself poised over just such a space. It seems extremely unlikely that the bacterium does so, even though it too displays allostatic responses and integrates bodily and worldly information as a means of determining next actions.

The gap, we suggested, is nothing other than a set of opportunities for control. It reflects the availability for control of action of information computed using a temporally deep generative model. Such a model needs to integrate bodily and exteroceptive information with goals and purposes at various timescales. When this occurs, there is the possibility for conflict between possible policies and courses of action. To be poised over an action space is thus to become informed of potentially conflicting possibilities in a way that invites further attempts at optimization – for example, stopping to reflect on what we are about to do.

Pezzulo et al (2018) note the important role of motivation in this process. We are not simply poised over an action space, nor are we simply poised over an action space that is allostatically inflected. Instead, we are poised over an action space in a way that makes contact with a complex, multi timescale set of motivations and priorities. To borrow their example, finding myself in a restaurant confronted with the dessert trolley, I encounter an action space ( a space of affordances) that is brought into contact with my own longer-term goals and wishes. According to their picture, a 'control hierarchy' (plausibly associated with the activities of dorsolateral PFC then exchanges messages with a 'motivational hierarchy' (plausibly associated with activities in ventromedial PFC), so as to drive action selection in a way sensitive to long-term goals and wishes, such as the wish to avoid sugary desserts – items that may be presenting undeniable short-term allostatic attractions. Motivations, reflecting both immediate context and longer term goals, alter the weightings (precisions) assigned to opportunities revealed by the control hierarchy. Motivated action occurs when high precision is assigned to one of the opportunities for action. At that moment, we are driven to act on our knowledge of what we can do. In this way, the revealed action space is placed in direct contact with affect, goals, and motivation, operating over temporally extended periods. The 'feel' of being poised to act reflects this combination of knowledge about control (what we can do) and knowledge about motivations and goals. In DF, the downstream impairment to areas crucial for visual form recognition restricts the kinds of information available for this kind of integration.

A particularly attractive element of this story is that it genalises to poise over mental action space as well. Mental action in the active inference framework builds straightforwardly on the planning as inference story. However, in this case the inference about hidden states concerns attentional states rather than the hidden states causing sensory impressions, and the state transitions refer to transitions between attentional states (Smith et al, 2020). Recall, in active inference, policy selection involves selecting the sequence of actions associated with the least expected free energy, based on beliefs on transitions between states. Inference about the hidden states themselves—perceptual inference or 'state estimation'—is based on a likelihood mapping that encodes beliefs about how the (hidden) states in the world relate to the observations they generate. Attentional processes are understood in terms of the 'precision'— the second order confidence in this likelihood mapping. In other words, the precision can be understood as an estimate of confidence that the agent has that their observations reliably map to hidden states in the world. In the same way that the agent can infer control over sensory inputs via action, and select policies accordingly, implicitly metacognitive modelling of attentional states means that the agent can perform covert actions in order to perform state transitions—move between—attentional states.

If this complex multi-dimensional story is on track, then experience emerges where (i) there is integrated bodily and worldly information computed using a generative model that displays temporal depth, and (ii) where that model integrates  control and motivation across many timescales, bringing goals and affect into direct contact with an appreciation of the space of possible actions that are currently enabled. When those twin conditions (resulting in a highly complex set of 'generative entanglements' - see Clark 2019) are met, a creature knows the value-inflected action space that is currently made available by its own perceptual contact with the world.

## 8. Consciousness Deflated

Does this explain consciousness itself? What is on offer is really just a kind of engineering blueprint for a being that would (if it was able to talk) say that it encounters a space of opportunities for action and cares about how it negotiates that space. Could this perhaps all occur 'in inner darkness', without the guiding light of qualitative experience at all?

A full answer would require a very different paper (see Clark, Friston & Wilkinson, 2019). But it is intriguing to speculate that these very same capacities, when present to enhanced degrees, explain the tendency of some advanced agents to behave in ways that express puzzlement about their own conscious experience and to infer the existence of those mysterious 'qualia'. This would be suggestive, at the very least. Perhaps it is the knowing poise over an action space that explains the attractions of the simple model according to which we are home to mysterious qualia intervening between perception and response?

Recall that the combination of the control and motivational hierarchies is itself dependent upon powerful capacities of precision variation. At any given moment, we harbor many preferences and goals and it is their varying precisions (reflecting a combination of their strategic value and their current attainability) that determine which ones get to guide action and choice. This precision-weighted integration reflects what we want to do and achieve, and enables us to make the best use of the currently revealed action space. The resulting functional organization enables us to explore counterfactual futures conditioned upon our own goals and actions. But this facility with counterfactual reasoning can, in advanced agents, lead to puzzling discoveries.

One such discovery concerns the unreliability of sensory evidence itself. Creatures like us possess a very unusual, but by no means magical, ability. We are able to force our own precision assignments so as to explore radical scenarios in which very high confidence in the way an action space appears to us is combined with radical variations in the true environment. For example, we can force our own precision-assignments so as to see that  "if this were a film-set, everything might look and sound just as it does". What we confront here looks to be a unique-to-humans extension of a much more basic capacity shared with many other animals. The basic capacity is to appreciate a space of possibilities in advance of taking actual action in the world. Creatures that act to minimize expected future prediction error relative to goals are already movers and shakers in just that kind of space.

But suppose that some of those creatures learnt ways to exert ever-more deliberate control over their own counterfactual explorations. In predictive processing terms, that would mean learning to exert even greater control over assignments of precision. Such a creature might, for example, forcibly assign—as a kind of imaginative thought experiment—high precision to an 'I am trapped in the Matrix' belief, so as to become aware that that belief would actually be fully consistent with very confident beliefs about the apparent shape of the action space made available by current sensory evidence. At that point we learn a surprising fact – Matrix world would be sensorily indistinguishable from the normal one. This dramatic outcome might also emerge if they were to fix the high-level story as 'being in a dream' or 'being tricked by an evil demon'.

We humans—perhaps in part courtesy of our experiences with language and the effects of complex cultural immersion—have that skill. It clearly confers huge cognitive benefits. But it also fuels the fires of a dualist picture in which experience suddenly seems special. After all, we were imaginatively able dramatically to vary the big picture ("perhaps this is all happening to me in the Matrix?") while keeping our sense of poise over an action space firmly fixed. Intelligent agents with that kind of command over their own counterfactual spaces can then begin to contemplate a very strange idea – one that leads us down many philosophical and scientific rabbit holes. They can realize that there might turn out to be nothing in the real world bearing sensory properties—of shape, colour, touch, and smell—that are nonetheless very confidently computed, hence presented as actual, by our prediction error minimizing brains. For such beings, Cartesian doubt—doubt about whether we are living in some kind of dream—becomes possible.

This is surely an important step on the rocky road to dualism. By adding enhanced layers of control to precision-weighted processing, a being becomes able to forcibly hold strange top-level pictures in while noticing that their own sensory evidence (their knowing poise over an action space) remains remarkably untouched by that manipulation. This opens up a gap between sensory best-guessing and reality, and invites the being to fill that gap with some kind of construct- something confidently known yet strangely divorced from how things really are. Such beings will be led to say and do the kinds of things that we say and do when we are expressing puzzlement about our own inner mental life.

Perhaps we are those creatures. Our amazing abilities of counterfactual imaginative variation lead us to believe that we are home to puzzling 'qualitative states'. The states (various best-guesses at how things are in the body and world, and what to do about them) are real but the puzzlement is chimerical. It is itself just one more inference or guess. Human experience is an inferential mountain, and qualia are themselves inferred on the basis of puzzling patterns in counterfactual space.

## 9. Conclusions

Marchi and Hohwy stopped short of claiming that intermediate-level flagging is identical with phenomenal consciousness, arguing only that it resolves the 'scope' question. We suggest that, by locating their considerations within the larger organizing framework of both allostasis and the action-space account, we make it plausible to assert a stronger claim.  What consciousness is, we suggest, is  nothing other than the process of inferring actionable policies, where that requires exploring possibilities defined at the intermediate level of the generative model – because this is the level at which  precise actionable policies can be optimally inferred. Creatures with very limited action repertoires have no need to flag or otherwise highlight such a level. Nor do creatures whose reactions to stimuli are all reflex-like, hence defined very close to the sensory stimulations themselves. But creatures whose generative models span many spatiotemporal scales, and whose projects and goals take many forms, are creatures that might otherwise find themselves constantly inferring policies that they cannot successfully implement. Flagging (if this is the right metaphor – nothing is meant to hang on this choice) should therefore develop as temporal depth in the generative model increases.

The question of why flagging this information should feel like anything at all looks less pressing once we realize that interoceptive predictions are constantly in play, turning the bare action space into an allostatically sensitive arena: an action space populated by opportunities to serve bedrock organismic needs. Intermediate level processing now reveals a world of changing opportunities that are often viscerally salient. Poised over these enriched action spaces, we are immersed in a world of possible actions and ineliminable mattering.

Still not enough? Beyond all this, much of our own puzzlement about our own conscious experience seems explicable by appeal to our highly advanced abilities of control, enabling us to explore strange counterfactual spaces featuring Cartesian demons and zombies. Intriguingly, it is the presence of that flagged arena of intermediate processing that, when combined with advanced control over our own precision-weighted processing, makes the hard puzzle seem both pressing and insoluble.

If nothing else, putting all these pieces together suggests that the active inference framework, suitably elaborated, has the resources to offer a principled account of the scope of conscious content, while revealing more about its profound links with allostasis, action and temporal depth.

# Conclusions and future directions

This thesis has taken an indirect route to an argument for consciousness in the active inference framework. Instead of addressing consciousness directly, I have focused on accounting for the content of consciousness, or how subjectivity is 'shaped' by phenomenal self-models. On this view, the experience of being a self is understood in terms of an inference about control of self-evidencing outcomes across multiple temporal scales. I have then argued that these self-modelling mechanisms are critical to shaping consciousness itself, forming the 'lens' of perception. On this view, experience is filtered through an inference about what does this sensation mean *for me,* as an embodied organism. In cases where these self-modeling mechanisms break down, such as in the psychedelic state, the basic facet of experience is shown to be affective.

Chapter 4 saw how, conceptually, the computational underpinning of phenomenal self-models can broadly be broken down into 'agentive control' and 'motivation', where agentive control can be understood in terms of an inference about the control of sensory inputs via action, and 'motivation' refers to the prior preferences that the system is biased to fulfil. While conceptually these are separable, the phenomenology of selfhood—that is, the sense of agency and affectivity—is understood here in terms of an inference about control of prior preferences. This inference serves to tune the organism to the affordance landscape to realise self-evidencing outcomes—a simple illustration of this being the hungry or hot lioness in being drawn to the opportunity for food or the shade under a tree respectively. In organisms with deep temporal models, this can be extended to temporally distal outcomes, such as a high arousal state in anticipation of an approaching deadline.

The phenomenology of being a self is one of the most pervasive aspects of experience, accompanying almost all typical states of consciousness. 'Subjectivity theories', equate consciousness with self-consciousness, or at least claim that self-consciousness is a necessary constituent of the conscious condition. To bring the relationship between consciousness and self-consciousness into focus, I have given special attention to the instances where computational self-modelling mechanisms and phenomenal self-modelling mechanisms come apart.

Instances where this phenomenology breaks down can be particularly illuminating in unpicking how ordinary experience is constructed, and serve to illuminate the relationships between consciousness, selfhood, agency and affectivity.

This thesis has highlighted the two primary aspects of consciousness *sensorimotor* and *affective*. In chapter 5, I argued that affectivity should be understood as the necessary constituent of consciousness. While the sensorimotor dimension of consciousness is not necessary, I have argued that this structures experience and underpins more complex affective states.

I have argued that understanding consciousness other species critically involves understanding the sensorimotor aspects of consciousness—the inference about control of sensation via action, as these counterfactually rich expectations—not only underpin our experience of a world of objects and events—but inference about control across hierarchical levels also shape the affective dimensions of consciousness. In chapter 3, I argue that in the psychedelic state, sensorimotor aspects of consciousness can break down—and with them the feeling of being an agent. In chapter 5, I argue that the fact that sensorimotor aspects of consciousness can breakdown provides reason to think that affectivity should be considered the constituent factor of phenomenal consciousness.

However, it is worth acknowledging that the presence of sensorimotor consciousness in the absence of affective consciousness cannot be ruled out. If sensorimotor consciousness and evaluative consciousness are indeed separable dimensions of consciousness, this raises the possibility of two types of phenomena that are grouped under "subjective experience". This is a point raised in Godfrey-Smith (2019):

> "If we ask, introspectively, about conspicuous features of human experience that may have early forms, it might be intuitive that one side of the phenomenon involves tracking external objects and events as external – achieving a point of view on things – while another involves distinctions between good and bad, a distinction that might be present in phenomenal washes that have no definite referral to organism or to environment." (p. 14)

Godfrey-Smith goes on to note that some spiders demonstrate complex perceptual capacities, but score low in respect to evidence for complex or varying motivational states. This would be expected in creatures that, given an evolutionary niche, don't need a more sophisticated "domain general controller" instantiated by affective inference. Other creatures—such as certain gastropods—may have richer subjective valuation in the absence of more complex sensorimotor capacities.

## The ontogeny of selfhood

Chapter 5 gives some discussion as to attribution of consciousness to non-human animals, identifying complex sensory attenuation mechanisms as the critical marker. We might also want to ask what this account means for the ontogeny of phenomenal selfhood. I have argued that the sense of self is underpinned by inferences about expected control of sensations are inferred through action and observation of action-outcome contingencies. Based on this view, there is evidence to suggest that the sense of self is learned over the course of ontogeny, in accordance with evidence suggesting that motor and brain development are closely intertwined with the process of self-exploration, where internal bodily representations are formed through learning of action-outcome contingencies (Cang & Feldheim, 2013).

Ma & Hommel (2015) provide a compelling a compelling illustration of how of how a bodily representation of the self could be formed over ontogeny. In their experiment, a virtual balloon or virtual square changes in size or colour reliably with certain movements of the participant's hand. They find that healthy adults can perceive body ownership illusions for the virtual objects, resulting from the congruence between predicted sensory outcomes and motor actions. For instance, when participants moved or rotated their (hidden) hand, the balloon would change, such as growing bigger or smaller with the hand opening and closing. Over time, the participants develop a sense of "ownership" of the virtual object, as if it was part of their body. This builds on earlier work showing that synchronous contingencies between actions and outcomes induced an increased sense perceived agency (and bodily ownership) than asynchronous contingencies (Ma & Hommel, 2013).

Similar processes may underlie how infants learn a sense of control over the course of development. The first sense to emerge in foetuses is thought to be the somatosensory sense (Bradley & Mistretta, 1975). Foetuses engage in self-touch in the womb, often gravitating towards the most innervated areas like the mouth and the feet, and moving onto other parts of the body, suggestive that foetuses are preferentially seeking movements that are informative (Fagard, Esseily, Jacquey, O'Regan, & Somogyi, 2018). There is evidence to suggest that as early as 19 weeks foetuses seem to have the rudiments of body representation, in that they are able to anticipate hand-to-mouth touch, as they open their mouths prior to contact (Myowa-Yamakoshi & Takeshita, 2006). Foetuses of 22 weeks show evidence of goal states associated with particular actions, as actions become more directed depending on the action goal (Zoia et al., 2007). When exactly a unified model comes online is not clear – for instance there is evidence suggesting that multimodal represention pertaining to higher levels are formed during the first year after birth (Hoffmann, 2017), where in the first weeks of life, "infants develop an ability to detect intermodal invariants and regularities in their sensorimotor experience, which specify themselves as separate entities agent in the environment." (Rochat, 1998, p. 1) More refined models of body representation appear to arise though exploratory behaviour of the body over time. Hoffman et al (2017) observed how infants reacted to a vibrotactile stimulus applied to different parts of the body over a period of 3-21 months. They found substantial development, particular over the course of the first year, as infant's ability to successfully reach for and remove the buzzer improved. In particular, during the first 3-4 months, response patterns were non-specific, where the infant might move their whole body, suggestive of exploratory motor activity, generating random motor babbling with self-touch occurring spontaneously. From 4-12 months the actions become more goal directed, and the infant responds with specific and direct movement to the location of the buzzer (Hoffmann et al., 2017). When infants reach about four months of age it has been shown that the timing of their smiles is goal-directed, where the action (smiling) has the outcome of causing the mother to smile (Ruvolo, Messinger, & Movellan, 2015; Blakemore, Frith, & Wolpert, 1999). Through a predictive processing lens, this can be understood as refining expectations of action-outcome contingencies, as the foetus learns to predict the sensory outcomes of motor actions on its body (Blakemore et al., 1999). Experiments such as these suggest that through self-exploration through motor babbling, the brain is learning the statistical connections between action outputs and sensory inputs, over time integrated this information into unified sensorimotor percepts and a unified self-model. While it is unclear where to draw a line as to when a unified self-model comes online from this evidence, it is suggestive that infants learn to posit themselves as a causally efficacious latent variable or 'endogenous cause'.

## Consciousness in Artificial systems

In principle, the there seems no reason to think the architecture of an allostatic control model couldn't be implemented in artificial systems. It's worth considering how artificial systems could adopt similar principles, such as a humanoid robot that can learn to expect outcomes from self-generated movements through self-exploration behaviour (Lang, Schillaci, & Hafner, 2018). Here, a deep convolutional neural network mapped proprioceptive (e.g. arm starting point) data and motor data (applied motor commands onto the expected visual outcomes of the actions). A forward model then computes a prediction error—the discrepancy actual visual inputs with expected visual inputs. The authors reason that a core component of a sense of agency is captured here, as prediction errors may serve as a cue for distinguishing between exogenous and endogenous causes of sensory signals (that is, distinguishing between self and other). Their model also captures sensory attenuation, a key component of self-generated action, as a corollary discharge cancels out expected sensory inputs (understood as a reduction of precision on sensory channels). In their experiment, sensory attenuation of self-generated sensory signals has the benefit of enhancing visual perception of objects that are occluded by the robot body. Similarly Schillaci et al (2016) present a biologically inspired model for learning multimodal body representations in artificial agents. They showed that coding internal body representations can generate predictions of auditory and motor inputs. There was greater sensory attenuation (e.g. a reduction in 'ego-noise') when the robot is the owner of the action, and greater prediction error resulted when the inputs were from a simulated robot body providing 'incongruent' inputs (Schillaci, Ritter, Hafner, & Lara, 2016).

These cases, in only implementing 'lower-order' sensory attenuation mechanisms, can be thought of as analogous to the case of the C.Elegans, unlikely to yield a self-model. In practice, the creation of artificial systems making use of something akin to the higher-order corollary discharge—which I've argued is likely to be indicative of a model of allosatic control—will perhaps involve an incremental process of gradually building up a model of motivations and abilities, in much the same way that it does in humans and non-human animals. This point echoes the developmental approach advocated by Alan Turing, who suspected it would be easier to build and educate a 'child' machine than it would to give a machine adult human level cognition (Turing, 1950). Lake et al (2016) suggest that a "developmental start-up software" may be key to the development of artificial general intelligence. Recent developments lay groundwork for this kind of approach, based on the sensorimotor experiences evoked by bodily movements beginning in the foetal period (Yamada et al., 2016). Through incorporating a cortex, spinal circuit and musculoskeletal body with sensory receptors for tactile perception, proprioception and vision, and a wealth of anatomical and physiological data, Yamada et al (2016) created a model was use to simulate cortical learning based on spontaneous bodily movements. It showed that "intrauterine sensorimotor experiences can facilitate the cortical learning of body representation and subsequent visual-somatosensory integration" (Yamada et al., 2016, p. 1).

Numerous questions are left outstanding for further theoretical and empirical investigation. Aligning with many accounts in the literature that posit embodiment as a critical feature of selfhood (e.g. Seth & Tsakiris, 2018), the account in this thesis leans heavily on modelling of the physiological condition of the body via interoception. The question remains, however, whether a body—rather than simply a means of closing an action-perception loop—is really necessary. It is not clear based on the present proposal whether a virtual body, or indeed just an inference over subjective fitness that informs action selection, would suffice to instantiate an experience, as it is clear that the allostatic control model can frequently operate in the absence of overt action, as in the case of mental action. It is conceivable that an agent could only care about improving its epistemic grip on the world (although, indeed, this might also involve pragmatic considerations for instrumental reasons). Whether or not such an agent would have subjectivity is an open question, and speaks to the question of whether an allostatic control model necessarily needs to be situated within something like the free energy principle at all.

# References

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

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. (2013). The computational anatomy of psychosis. Frontiers in Psychiatry, 4, 47.

Ahrens, M. B., Li, J. M., Orger, M. B., Robson, D. N., Schier, A. F., Engert, F., & Portugues, R. (2012). Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature*, *485*(7399), 471–477.

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

Aitchison, L., & Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227.

Alamia, A., Timmermann, C., Nutt, D. J., VanRullen, R., & Carhart-Harris, R. L. (2020). DMT alters cortical travelling waves. *Elife*, *9*, e59784.

Albahari, M. (2009). Witness-consciousness: Its definition, Appearance and reality. *Journal of Consciousness Studies*, *16*(1), 62-84.

Albahari, M. (2014). Insight knowledge of no self in Buddhism: An epistemic analysis. *Philosopher's Imprint*, *14*(21).

Allen, M., & Friston, K. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. Synthese, 195(6), 2459–2482.

Allen, M., & Tsakiris, M. (2018). The body as first prior: Interoceptive predictive processing and the primacy. In *The Interoceptive Mind: From Homeostasis to Awareness* (pp. 27–45).

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K. J., Rees, G., & Roepstorff, A. (2016). Anterior insula coordinates hierarchical processing of tactile mismatch responses. *Neuroimage*, *127*, 34-43

Aminoff, E. M., & Tarr, M. J. (2015). Associative processing is inherent in scene perception. *PLoS One*, *10*(6), e0128840.

Anālayo. (2009). *From Craving to Liberation-Excursions Into the Thought-word of the Pāli Discourses (1).* Buddhist Association of the United States.

Anchisi, D., & Zanon, M. (2015). A Bayesian perspective on sensory and cognitive integration in pain perception and placebo analgesia. PloS One, 10(2), e0117270.

Anderson, E., Siegel, E., White, D., & Barrett, L. F. (2012). Out of sight but not out of mind: unseen affective faces influence evaluations and social impressions. *Emotion*, *12*(6), 1210.

Anderson, M. L., (2017) Of Bayes and bullets: an embodied, situated, targeting-based account of predictive processing. In *Philosophy and predictive processing: 3* (eds Metzinger T, Wiese W), pp. 60–73. Frankfurt am Main, Germany: MIND Group.

Andrews, M. (2020). The Math is not the Territory: Navigating the Free Energy Principle.

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

Artyukhin, A. B., Yim, J. J., Cheong, M. C., & Avery, L. (2015). Starvation-induced collective behavior in C. elegans. *Scientific Reports*, *5*(1), 1–10.

Ashby, W. (2013). *Design for a brain: The origin of adaptive behaviour*. Springer Science & Business Media.

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

Atlas, L. Y., & Wager, T. D. (2012). How expectations shape pain. *Neuroscience Letters*, *520*(2), 140–148.

Atlas, L. Y., Lindquist, M. A., Bolger, N., & Wager, T. D. (2014). Brain mediators of the effects of noxious heat on pain. *PAIN®*, *155*(8), 1632–1648.

Attias, H. (2003). Planning by probabilistic inference. In *AISTATS*. Citeseer.

Austin, J. H. (2013). Zen and the brain: mutually illuminating topics. *Frontiers in psychology*, *4*, 784.

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The depressed brain: an evolutionary systems theory. *Trends in Cognitive Sciences*, *21*(3), 182-194.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., & Kauffman, S. (2019). The hierarchically mechanistic mind : A free-energy formulation of the human psyche. *Physics of Life Reviews*, *1*, 1–18.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020, July). Predictions in the eye of the beholder: an active inference account of Watt governors. In *Artificial Life Conference Proceedings* (pp. 121-129). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617-629.

Barlow, H. (2001). Redundancy reduction revisited. Network: Computation in Neural Systems, 12(3), 241–253.

Barre, A., Berthoux, C., De Bundel, D., Valjent, E., Bockaert, J., Marin, P., & Bécamel, C. (2016). Presynaptic serotonin 2A receptors modulate thalamocortical plasticity and associative learning. Proceedings ofthe National Academy of Sciences, 113(10), E1382–E1391.

Barrett, L. F. (2014). The Conceptual Act Theory: A Precis. *Emotion Review* 6 (4):292-297.

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

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419.;

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

Barrett, L.F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10, 20-46.

Barry, T. J., Vervliet, B., & Hermans, D. (2015). An integrative review of attention biases and their contribution to treatment for anxiety disorders. *Frontiers in psychology*, *6*, 968.

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition* 11:211-277.

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems* (pp. 17-47). Springer, Berlin, Heidelberg

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise?. *Frontiers in psychology*, *4*, 907.

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

Bayne, T., & Carter, O. (2018). Dimensions of consciousness and the psychedelic state. Neuroscience of Consciousness, 2018(1), niy008.

Beliveau, V., Ganz, M., Feng, L., Ozenne, B., Højgaard, L., Fisher, P. M., et al. (2017). A high-resolution in vivo atlas of the human brain's serotonin system. Journal ofNeuroscience, 37(1), 120–128.

Bell, C. C. (1981). An efference copy which is modified by reafferent input. *Science*, *214*(4519), 450–453.

Benedetti, F., Carlino, E., & Pollo, A. (2011). Hidden administration of drugs. *Clinical Pharmacology & Therapeutics*, *90*(5), 651–661.

Benedetti, F., Maggi, G., Lopiano, L., Lanotte, M., Rainero, I., Vighetti, S., & Pollo, A. (2003). Open versus hidden medical treatments: The patient's knowledge about a therapy affects the therapy outcome. *Prevention & Treatment*, *6*(1), 1a.

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* (Berl.) 191: 391–431.

Berthoux, C., Barre, A., Bockaert, J., Marin, P., & Bécamel, C. (2018). Sustained activation of postsynaptic 5-HT2A receptors gates plasticity at prefrontal cortex synapses. Cerebral Cortex, 29(4), 1659–1669.

Billon, A., & Kriegel, U. (2016). Jaspers' Dilemma: The Psychopathological Challenge to Subjectivity Theories of Consciousness. *Disturbed Consciousness*, 29–54.

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of cognitive neuroscience*, *11*(5), 551-559.

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature neuroscience*, *1*(7), 635-640.

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

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. Trends in Cognitive Sciences, 13(1), 7–13.

Botvinick, M., & Cohen, J. (1998). Rubber hands "feel" touch that eyes see [8]. *Nature*, *391*(6669), 756.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in cognitive sciences*, *16*(10), 485-488.

Bowen, S., & Marlatt, A. (2009). Surfing the urge: brief mindfulness-based intervention for college student smokers. Psychology of Addictive Behaviors, 23(4), 666.

Bradley, R. M., & Mistretta, C. M. (1975). Fetal sensory receptors. Physiological Reviews, 55(3), 352–382.

Brainard, M. S., & Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, *1*(1), 31.

Bridgeman, B., Kirch, M., & Sperling, A. (1981). Segregation of cognitive and motor aspects of visual function using induced motion. *Perception & Psychophysics*, *29*(4), 336-342.

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

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

Bruineberg, J. (2017). Active inference and the primacy of the 'I can'. In T. Metzinger & W. Wiese (Eds.) Philosophy and predictive processing. Frankfurt am Main: MIND Group

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in human neuroscience*, *8*, 599.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417-2444.

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, *4*, 25.

Büchel, C., Geuter, S., Sprenger, C., & Eippert, F. (2014). Placebo analgesia: A predictive coding perspective. Neuron, 81(6), 1223–1239.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55-79.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2), 49–57.

Cabanac, M. (1992). Pleasure: the common currency. *Journal of Theoretical Biology*, *155*(2), 173–200.

Calvo, P., & Friston, K. (2017). Predicting green: really radical (plant) predictive processing. *Journal of the Royal Society Interface*, *14*(131), 20170096.

Campbell, J. O. (2016). Universal Darwinism as a process of Bayesian inference. *Frontiers in Systems Neuroscience*, *10*, 49.

Cang, J., & Feldheim, D. A. (2013). Developmental mechanisms of topographic map formation and alignment. Annual Review of Neuroscience, 36, 51–77.

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, *39*(1/4), 106-124

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, 9(3), 399–431.

Carbonaro, T. M., Johnson, M. W., Hurwitz, E., & Griffiths, R. R. (2018). Double-blind comparison of the two hallucinogens psilocybin and dextromethorphan: similarities and differences in subjective experiences. *Psychopharmacology*, *235*(2), 521–534.

Carhart-Harris, R. L. (2019). How do psychedelics work? Current Opinion in Psychiatry, 32(1), 16–21.

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

Carhart-Harris, R. L., & Nutt, D. (2017). Serotonin and brain function: A tale of two receptors. *Journal of Psychopharmacology*, 31(9), 1091–1120.

Carhart-Harris, R. L., Roseman, L., Haijen, E., Erritzoe, D., Watts, R., Branchi, I., & Kaelen, M. (2018). Psychedelics and the essential importance of context. *Journal of Psychopharmacology*, 32(7), 725–731.

Carter, O. L., Hasler, F., Pettigrew, J. D., Wallis, G. M., Liu, G. B., & Vollenweider, F. X. (2007). Psilocybin links binocular rivalry switch rate to attention and subjective arousal levels in humans. *Psychopharmacology*, 195(3), 415–424.

Carter, O. L., Pettigrew, J. D., Hasler, F., Wallis, G. M., Liu, G. B., Hell, D., & Vollenweider, F. X. (2005). Modulating the rate and rhythmicity of perceptual rivalry alternations with the mixed 5-HT 2A and 5-HT 1A agonist psilocybin. *Neuropsychopharmacology*, 30(6), 1154.

Carter, O. L., Presti, D. E., Callistemon, C., Ungerer, Y., Liu, G. B., & Pettigrew, J. D. (2005). Meditation alters perceptual rivalry in Tibetan Buddhist monks. *Current Biology*, *15*(11), R412-R413.

Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological review*, 97(1), 19.

Castiello, U., Paulignan, Y., & Jeannerod, M. (1991). Temporal dissociation of motor responses and subjective awareness: A study in normal subjects. *Brain*, *114*(6), 2639-2655

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1994). A critique of pure vision. *Large-scale neuronal theories of the brain*, *23*

Ciaunica, A., Charlton, J., & Farmer, H. When the Window Cracks: Transparency and the Fractured Self in Depersonalisation. *Phenomenology and the Cognitive Sciences*, 1-19.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485), 1585–1599.

Clark-Polner, E., Johnson, T. D., and Barrett, L. F. (2016). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cereb. Cortex* 27, 1944–1948. Clark, A. (1997). *Being there* (p. 222). Cambridge, MA: MIT Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA

Clark, A. (2012). Dreaming the Whole Cat: Generative Models. *Predictive Processing, and Enactivist.*

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

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

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

Clark, A. (2017). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*, *51*(4), 727-753

Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, 116, 645–662.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7-19

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, *26*(9–10), 19–33.

Clark, J. E., Watson, S., & Friston, K. J. (2018). What is mood? A computational perspective. *Psychological Medicine*, *48*(14), 2277-2284

Colombetti, G. (2014). *The feeling body: Affective science meets the enactive mind.* MIT press:

Colombetti, G., & Ratcliffe, M. (2012). Bodily feeling in depersonalization: A phenomenological account. *Emotion Review*, *4*(2), 145-150.

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

Corcoran, A. W., & Hohwy, J. (2017). Allostasis , interoception , and the free energy principle : Feeling our way forward. *Interoceptive Basis of the Mind.*

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 1–45.

Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4), 515–530.

Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology*, 30(11), 1145–1155.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and Strong Priors. *Trends in Cognitive Sciences*, *23*(2), 114–127.

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature reviews neuroscience*, *3*(8), 655

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, *13*(4), 500-505

Craig, A. D., & Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. *Nature reviews neuroscience*, *10*(1)

Crapse, T. B., & Sommer, M. A. (2008a). Corollary discharge across the animal kingdom. Nature Reviews Neuroscience, 9(8), 587.

Crapse, T. B., & Sommer, M. A. (2008b). Corollary discharge circuits in the primate brain. *Current Opinion in Neurobiology*, *18*(6), 552–557.

Crick, F. (1994). The astonishing hypothesis: The science search for the soul. *New York: Touchstone.*

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective and cognitive integration. *J. Comp. Neurol.* 493, 154–166.

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current opinion in psychology*, *17*, 7-14.

Critchley, H. D., & Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron*, *77*(4), 624-638

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature neuroscience*, *7*(2), 189.

Damasio, A. (2003). Feelings of emotion and the self. *Annals of the New York Academy of Sciences*, *1001*(1), 253-261.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Damasio, A. R. (2019). *The strange order of things: Life, feeling, and the making of cultures*. Vintage.

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, 14(2), 143.

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.

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.

Danckert, J., & Rossetti, Y. (2005). Blindsight in action: what can the different sub-types of blindsight tell us about the control of visually guided actions?. Neuroscience & Biobehavioral Reviews, 29(7), 1035-1046.

David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*, *17*(2), 523–534.

Davis, J., & Thompson, E. (2017). From the five aggregates to phenomenal consciousness: toward a cross-cultural cognitive science.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429-453.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, *7*(5), 889-904.

Deane, G. (2020). Dissolving the self: Active inference, psychedelics, and ego-dissolution. *Philosophy and the Mind Sciences*, 1(2).

Deane, G., Miller, M. D., & Wilkinson, S. (2020). Losing Ourselves: Active Inference, Depersonalization and Meditation. *Frontiers in Psychology*, *11*, 2893.

Dehaene, S., Lau, H., Kouider, S., Silver, D., Huang, A., Maddison, C. J., … Donchin, E. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 484–489.

Denève, S. (2008). Bayesian spiking neurons I: inference. *Neural Comput.* 20, 91–117.

Dennett, D. C. (1993). *Consciousness explained.* Penguin uk.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the cognitive sciences*, *4*(4), 429-452

Doerig, A., Schurger, A., & Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *00*(00), 1–22.

Dupuis, D. (2020). The socialization of hallucinations. Cultural priors, social interactions and contextual factors in the use of ayahuasca. *Transcultural Psychiatry*, (August), 1–26.

Eisenberger, N. I. (2012). The pain of social disconnection: Examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience*, 13(6), 421.

Eisenberger, N. I., Jarcho, J. M., Lieberman, M. D., & Naliboff, B. D. (2006). An experimental study of shared sensitivity to physical pain and social rejection. Pain, 126(1-3), 132–138.

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24.

Eldar, S., Ricon, T., & Bar-Haim, Y. (2008). Plasticity in attention: Implications for stress response in children. *Behaviour Research and Therapy*, *46*(4), 450-461.

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature*, *453*(7198), 1102–1106.

Engel, A. K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in cognitive sciences*, *17*(5), 202-209.

Ernst, M. O., & Banks, M. S. (2002). Integrate Visual and Haptic_2002.Pdf, *415*(January), 429–433.

Fabry, R (2017). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, 1–20.

Fabry, R. E. (2019). Into the dark room: a predictive processing account of major depressive disorder. *Phenomenology and the Cognitive Sciences*, 1-20

Fagard, J., Esseily, R., Jacquey, L., O'Regan, K., & Somogyi, E. (2018). Fetal origin of sensorimotor behavior. Frontiers in Neurorobotics, 12, 23.

Fanciullacci, M., Bene, E. D., Franchi, G., & Sicuteri, F. (1977). Phantom limb pain: Sub-hallucinogenic treatment with lysergic acid diethylamide (LSD-25). Headache: The Journal ofHead and Face Pain, 17(3), 118–119.

Farb N, Daubenmier J, Price CJ, Gard T, Kerr C, Dunn BD, Klein AC, Paulus MP, Mehling WE: Interoception, contemplative practice, and health. *Front Psychol* 2015, 6

Feinberg, I. (1978). Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophrenia Bulletin*, *4*(4), 636.

Feldman, A. G., & Levin, M. F. (1995). The origin and use of positional frames of reference in motor control. *Behavioral and Brain Sciences*, *18*(4), 723–744.

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4, 215.

Feldman, H., & Friston, K. J. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, *4*(December), 1–23.

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4:215.

Fink, P. W., Foo, P. S., & Warren, W. H. (2009). Catching fly balls in virtual reality: A critical test of the outfielder problem. *Journal of vision*, *9*(13), 14-14

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

Ford, J. M., & Mathalon, D. H. (2005). Corollary discharge dysfunction in schizophrenia: Can it explain auditory hallucinations? International Journal ofPsychophysiology, 58(2-3), 179–189.

Ford, J. M., & Mathalon, D. H. (2005). Corollary discharge dysfunction in schizophrenia: can it explain auditory hallucinations? *International Journal of Psychophysiology*, *58*(2–3), 179–189.

Ford, J. M., Gray, M., Faustman, W. O., Roach, B. J., & Mathalon, D. H. (2007). Dissecting corollary discharge dysfunction in schizophrenia. Psychophysiology, 44(4), 522–529.

Ford, J. M., Gray, M., Faustman, W. O., Roach, B. J., & Mathalon, D. H. (2007). Dissecting corollary discharge dysfunction in schizophrenia. *Psychophysiology*, *44*(4), 522–529.

Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis*, *19*(1), 3–28.

Friston K, Shiner T, FitzGerald T, Galea JM, Adams R, et al. (2012) Dopamine, Affordance and Active Inference. *PLoS Comput Biol* 8(1): e1002327

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 815–836.

Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, *13*(7), 293-301

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

Friston, K. (2012a). Prediction, perception and agency. *International Journal of Psychophysiology*, *83*(2), 248-252

Friston, K. (2012b). Embodied inference and spatial cognition. Cognitive Processing, 13(1), 171–177.

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

Friston, K. (2019). A free energy principle for a particular physics. *ArXiv Preprint ArXiv:1906.10184*.

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

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural computation*, 29(10), 2633-2683.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, *77*(April), 388–402.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148-158.

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516.

Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. Biological Cybernetics, 102(3), 227–260.

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

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68, 862-879

Friston, K., Pezzulo, G., Cartoni, E., & Rigoli, F. (2016). Active Inference , epistemic value , and vicarious trial and error, *2*(2013), 322–339.

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., Rosch, R., Parr, T., Price, C., & Bowman, H. (2018). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 90, 486–501.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, 7, 598.

Friston, K., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, 8(1), e1002327.

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Front. Psychol.* 3:130.

Frith, C. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition*, *14*(4), 752–770.

Frith, C. D. (1987). The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action. *Psychological medicine*, *17*(3), 631-648.

Frith, C. D. (2014). *The cognitive neuropsychology of schizophrenia*. Psychology press.

Frith, U. (2003). Autism: Explaining the enigma. Oxford: Blackwell Publishing.

Gallagher, S. (2004). Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology*, *37*(1), 8–19.

Gallagher, S. (2010). Defining consciousness: The importance of non-reflective self-awareness.

*Pragmatics & Cognition*, *18*(3), 561–569.

Gallagher, S. (2013). *The Phenomenological Mind. The Phenomenological Mind.*

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind.* Oxford University Press

Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, *195*(6), 2627-2648

Garfield, J. L. (2015). Buddhism and modernity. *The Buddhist World*, 294.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. (2009). The mismatch negativity: A review of underlying mechanisms. Clinical Neurophysiology, 120(3), 453–463.

Gerken, L., Balcomb, F. K., & Minton, J. L. (2011). Infants avoid 'labouring in vain'by attending more to learnable than unlearnable linguistic patterns. *Developmental science*, 14(5), 972-979.

Gerrans, P. (2019). Depersonalization disorder, affective processing and predictive coding. *Review of Philosophy and Psychology*, *10*(2), 401-418.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *Neurons, Behavior, Data Analysis, and Theory*, 4(1), 1–10.

Gilead, M., Trope, Y., & Liberman, N. (2019). Above and Beyond the Concrete: The Diverse Representational Substrates of the Predictive Brain. *Behavioral and Brain Sciences*, (July), 1–63.

Girn, M., & Christoff, K. (2018). Journal of Consciousness Studies, 25(11-12), 131–154.

Girn, M., Roseman, L., Bernhardt, B., Smallwood, J., Carhart-Harris, R. L., & Spreng, N. (2020). LSD flattens the functional hierarchy of the human brain. *BioRxiv.*

Gładziejewski, P. (2015). Explaining mental phenomena with internal representations. A mechanistic perspective.

Godfrey-Smith, P. (2019). Evolving across the explanatory gap. *Philosophy, Theory, and Practice in Biology*, *11*.

Goo, E., Majerczyk, C. D., An, J. H., Chandler, J. R., Seo, Y. S., Ham, H., ... & Hwang, I. (2012). Bacterial quorum sensing, cooperativity, and anticipation of stationary-phase stress. *Proceedings of the National Academy of Sciences*, *109*(48), 19775-19780.

Goodale, M. A., Pelisson, D., & Prablanc, C. (1986). Large adjustments in visually guided reaching do not depend on vision of the hand or perception of target displacement. *Nature*, *320*(6064), 748-750.;

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1358), 1121-1127.

Griffiths, R. R., Johnson, M. W., Carducci, M. A., Umbricht, A., Richards, W. A., Richards, B. D., et al. (2016). Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: A randomized double-blind trial. Journal of Psychopharmacology, 30(12), 1181–1197.

Grof, S. (1980). *LSD psychotherapy*. Pomona, CA: Hunter House.

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

Guillot, M. (2017). I me mine: on a confusion concerning the subjective character of experience. *Review of Philosophy and Psychology*, *8*(1), 23–53.

Hafner, V. V., Loviken, P., Villalpando, A. P., & Schillaci, G. (2020). Prerequisites for an artificial self. *Frontiers in neurorobotics*, *14*.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, *18*(4), 196.

Hamilton-Blyth, S. (2013). *Early Buddhism: A new approach: The I of the beholder*. Routledge.

Harris, S. E. (2018). A Nirvana that Is Burning in Hell: Pain and Flourishing in Mahayana Buddhist Moral Thought. *Sophia*, *57*(2), 337-347.

Harrison, N. A., Gray, M. A., Gianaros, P. J., and Critchley, H. D. (2010). The embodiment of emotional feelings in the brain. *J. Neurosci.* 30, 12878–12884.

Hechler, T., Endres, D., & Thorwart, A. (2016). Why harmless sensations might hurt in individuals with chronic pain: About heightened prediction and perception of pain in the mind. *Frontiers in Psychology*, *7*(OCT), 1–7.

Heinks-Maldonado, T. H., Mathalon, D. H., Houde, J. F., Gray, M., Faustman, W. O., & Ford, J. M. (2007). Relationship of imprecise corollary discharge in schizophrenia to auditory hallucinations. Archives of General Psychiatry, 64(3), 286–296.

Helmholtz, H.L. (1867/1910). Handbuch der physiologischen Optik. Leipzig: L. Voss. Reprinted, with extensive commentary, in A. Gullstrand, J. von Kries & W. Nagel (Eds.) *Handbuch der physiologischen Optik* (3rd edn.).

Hendricks, P. S., Thorne, C. B., Clark, C. B., Coombs, D. W., & Johnson, M. W. (2015). Classic psychedelic use is associated with reduced psychological distress and suicidality in the united states adult population. Journal of Psychopharmacology, 29(3), 280–288.

Herzog, R., Mediano, P. A. M., Rosas, F. E., Carhart-Harris, R., Perl, Y. S., Tagliazucchi, E., & Cofre, R. (2020). A mechanistic model of the neural entropy increase elicited by psychedelic drugs. *Scientific Reports*, *10*(1), 1–12.

Hesp, C., Smith, R., Allen, M., Friston, K., & Ramstead, M. (2019). Deeply felt affect: The emergence of valence in deep active inference. PsyArXiv.

Hesp, C., Smith, R., Allen, M., Friston, K., & Ramstead, M. (2019). Deeply felt affect: the emergence of valence in deep active inference. PsyArXiv Preprints [Preprint]. Available at: https://psyarxiv.com/62pfd/(Accessed December 3, 2019).

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation*, *33*(1), 1–49.

Hesp, C., Tschantz, A., Millidge, B., Ramstead, M., Friston, K., & Smith, R. (2020). Sophisticated affective inference: Simulating anticipatory affective dynamics of imagining future events. *Communications in Computer and Information Science*, *1326*(October), 179–186.

Hoffmann, M. (2017). The role of self-touch experience in the formation of the self. *arXiv preprint arXiv:1712.07843*.

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche*, 13(1), 1–20.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3, 96.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology*, *3*, 96.

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

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

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition*, 47, 75–85.

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition*, *47*, 75–85.

Hohwy, J., & Michael, J. (2017). Why should any body have a self? In F. de Vignemont & A. Alsmith (Eds.), *The subject's matter: Self-consciousness and the body* (pp. 363–391). Cambridge, MA: MIT Press.

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness, (Reardon).

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701.

Holst, E., & Mittelstaedt, H. (1950). Das reafferenzprinzip. *Naturwissenschaften*, *37*(20), 464–476.

Holton, R. 2016. Review of Surfing Uncertainty: Prediction, Action, and the Embodied Mind, *Times Literary Supplement* October 7, 10-11

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71-77.

Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, *129*(1), 3-40

Hurley, S. L. (1998). *Consciousness in action*. Harvard University Press

Huxley, A. (1952). The Doors of perception. *Mental*, *98*, 2–24.

Huxley, A. (2010). The doors of perception: And heaven and hell. New York, NY: Random House.

Inzlicht, M., Bartholow, B. D., & Hirsh, J. B. (2015). Emotional foundations of cognitive control. *Trends in Cognitive Sciences*, *19*(3), 126–132.

Jakab, R. L., & Goldman-Rakic, P. S. (1998). *Proceedings of the National Academy of Sciences*, 95(2), 735–740.

James, W. (1961). *The varieties of religious experience* (p. 7291411). Amazon Distribution GmbH.

James, W. (1983). What Is an Emotion?[1884]. *Collected Essays and Reviews*, 244-275.

Jobst, B. M., Atasoy, S., Ponce-Alvarez, A., Sanjuán, A., Roseman, L., Kaelen, M., … Deco, G. (n.d.). Increased sensitivity to strong perturbations in a whole-brain model of LSD. *BioRxiv*, 2001–2021.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS computational biology*, *9*(6), e1003094;

Kant, I. (1781/1929) *Critique of pure reason,* trans. N. Kemp Smith. Macmillan.

Kaplan, F., & Oudeyer, P. Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in neuroscience*, *1*, 17

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

Karl, F. (2012). A free energy principle for biological systems. *Entropy*, *14*(11), 2100–2121.

Keller, G. B., & Hahnloser, R. H. R. (2009). Neural processing of auditory feedback during vocal practice in a songbird. *Nature*, *457*(7226), 187–190.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. Neuron, 100(2), 424–435.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PloS one, 7(5), e36399.

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

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain, *4*(11).

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387-2415

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, *8*(3), 159-166

King, A., Martin, I., & Seymour, K. (1972). Reversal learning facilitated by a single injection of lysergic acid diethylamide (LSD 25) in the rat. British Journal ofPharmacology, 45(1), 161P.

Kira, I. A., Ashby, J. S., Odenat, L., & Lewandowsky, L. (2013). The mental health effects of torture trauma and its severity: A replication and extension. *Psychology, 4*(5), 472–482

Kirchhoff, M. & Kiverstein, J.,(2019*). Extended Consciousness and Predictive Processing : A Third-Wave View*

Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, *195*(6), 2519-2540

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: a non-representational view. *Philosophical Explorations*, *21*(2), 264-281.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, *15*(138), 20170792

Kirsch, I., Kong, J., Sadler, P., Spaeth, R., Cook, A., Kaptchuk, T. J., & Gollub, R. (2014). Expectancy and conditioning in placebo analgesia: separate or connected processes? *Psychology of Consciousness: Theory, Research, and Practice*, *1*(1), 51.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, *18*(4), 513-549.

Kiverstein, J., Miller, M., & Rietveld, E. (2017). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese*, *196*(7), 2847-2869.

Kiverstein, J., Miller, M., & Rietveld, E. (2020). How mood tunes prediction: a neurophenomenological account of mood and its disturbance in major depression. *Neuroscience of Consciousness*, *2020*(1), niaa003.

Klein, C. (2016). What do predictive coders want? *Synthese.*

Klein, C., & Barron, A. (2016). Insect consciousness : commitments, conflicts and consequences.

*Animal Sentience*, *1*(9), 1–12.

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

Koban, L., & Wager, T. D. (2016). Beyond conformity: Social influences on pain reports and physiology. *Emotion*, *16*(1), 24.

Kolodny, O., Moyal, R., & Edelman, S. (2021). A possible evolutionary function of pain and other affective dimensions of phenomenal conscious experience.

Kometer, M., Cahn, B. R., Andel, D., Carter, O. L., & Vollenweider, F. X. (2011). The 5-HT2A/1A agonist psilocybin disrupts modal object completion associated with visual hallucinations. *Biological Psychiatry*, 69(5), 399–406.

Kong, J., & Benedetti, F. (2014). Placebo and nocebo effects: an introduction to psychological and biological mechanisms. In *Placebo* (pp. 3–15). Springer.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. Behavioral and Brain Sciences, (2012), 1–101.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Lang, C., Schillaci, G., & Hafner, V. V. (2018, September). A deep convolutional neural network model for sense of agency and object permanence in robots. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 257-262). IEEE.

Laukkonen, R., & Slagter, H. A. (2020) From many to one: Meditation and the plasticity of the predictive mind.

Leary, T. F., Metzner, R., & Alpert, R. (1964). The psychedelic experience: A manual based on the Tibetan book of the dead.

Leary, T. F., Metzner, R., & Alpert, R. (1964). The psychedelic experience: A manual based on the Tibetan book of the dead.

Lebedev, A. V, Lövdén, M., Rosenthal, G., Feilding, A., Nutt, D. J., & Carhart-Harris, R. L. (2015). Finding the self by losing the self: Neural correlates of ego-dissolution under psilocybin. *Human Brain Mapping*, *36*(8), 3137–3153.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434-1448

Letheby, C. (2020). Being for no-one: Psychedelic experience and minimal subjectivity. *Philosophy and the Mind Sciences*, 1(I), 5.

Letheby, C., & Gerrans, P. (2017). Self unbound: Ego dissolution in psychedelic experience. Neuroscience of Consciousness, 3(1), nix016.

Limanowski, J. (2017). (Dis-) attending to the Body: Action and Self-experience in the Active Inference Framework.

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, *7*, 547.

Limanowski, J., & Friston, K. (2018). 'Seeing the dark': grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in Psychology*, *9*, 643.

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

Lindahl, J. R., & Britton, W. B. (2019). 'I Have This Feeling of Not Really Being Here': Buddhist Meditation and Changes in Sense of Self. *Journal of Consciousness Studies*, 26(7-8), 157-183.

Lofthouse, G., (2014). Enlightenment's Evil Twin, retrieved 25 February 2020 from http://www.theatlantic.com/health/archive/2014/12/enlightenments-eviltwin/383726/

Lou, H. C., Changeux, J. P., & Rosenstand, A. (2017). Towards a cognitive neuroscience of self-awareness. *Neuroscience & Biobehavioral Reviews*, *83*, 765–773.

Lüersen, K., Faust, U., Gottschling, D.-C., & Döring, F. (2014). Gait-specific adaptation of locomotor activity in response to dietary restriction in Caenorhabditis elegans. *Journal of Experimental Biology*, *217*(14), 2480–2488.

Lukitsch, O. (2020). Effort, Uncertainty, and the Sense of Agency. *Review of Philosophy and Psychology*, *11*, 955-975

Lunghi, C., & Morrone, M. C. (2013). Early interaction between vision and touch during binocular rivalry. *Multisensory Research*, *26*(3), 291–306.

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.

Ly, C., Greb, A. C., Cameron, L. P., Wong, J. M., Barragan, E. V., Wilson, P. C., et al. (2018). Psychedelics promote structural and functional neural plasticity. *Cell Reports*, 23(11), 3170–3182.

Lyons, T., & Carhart-Harris, R. L. (2018). More realistic forecasting of future life events after psilocybin for treatment-resistant depression. *Frontiers in Psychology*, 9, 1721.

Ma, K., & Hommel, B. (2013). The virtual-hand illusion: effects of impact and threat on perceived ownership and affective resonance. *Frontiers in Psychology*, 4, 604. 39

Ma, K., & Hommel, B. (2015). Body-ownership for actively operated non-corporeal objects. *Consciousness and Cognition*, 36, 75–86.

MacLeod, C., Rutherford, E., Campbell, L., Ebsworthy, G., & Holker, L. (2002). Selective attention and emotional vulnerability: assessing the causal basis of their association through the experimental manipulation of attentional bias. *Journal of abnormal psychology*, *111*(1), 107.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, *1*(10), 446-452.

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

Marchi, F., & Hohwy, J. (2020). The intermediate scope of consciousness in the predictive mind. *Erkenntnis*, 1-22.

Margoliash, D. (2002). Evaluating theories of bird song learning: implications for future directions. *Journal of Comparative Physiology A*, *188*(11–12), 851–866.

Marks, J. (1982). A theory of emotion. *Philosophical Studies* 42 (1):227-242.

Marr, D. (1982). Vision: A computational approach. San Francisco, CA: Freeman & Co.

Masters, R. E., & Houston, J. (1966). The varieties of psychedelic experience (Vol. 9289). New York, Chicago, San Francisco: Holt, Rinehart; Winston New York.

Mathys, C., Daunizeau, J., Friston, K., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5, 39.

Mathys, C., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K., & Stephan, K. E. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in Human Neuroscience*, 8, 825

Matthews, G. A., & Tye, K. M. (2019). Neural mechanisms of social homeostasis. *Annals of the New York Academy of Sciences*, *1457*(1), 5-25.

McNally, R. J. (2019). Attentional bias for threat: Crisis or opportunity?. *Clinical psychology review*, *69*, 4-13.

Medford, N. (2012). Emotion and the unreal self: depersonalization disorder and de-affectualization. *Emotion Review*, *4*(2), 139-144

Medford, N., Brierley, B., Brammer, M., Bullmore, E. T., David, A. S., & Phillips, M. L. (2006). Emotional memory in depersonalization disorder: a functional MRI study. *Psychiatry Research: Neuroimaging*, *148*(2-3), 93-102.

Medford, N., Sierra, M., Stringaris, A., Giampietro, V., Brammer, M. J., & David, A. S. (2016). Emotional experience and awareness of self: functional MRI studies of depersonalization disorder. *Frontiers in psychology*, *7*, 432

Menary, R. (2008). *Embodied narratives.*

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, *30*(1), 63–81.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

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

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self.* New York: Basic Books.

Metzinger, T. (2016). Suffering. In *The Return of Consciousness A new science on old questions*. Eds. Kurt Almqvist & Anders Haag. © Axel and Margaret Ax:son Johnson Foundation and the authors.

Metzinger, T. (2017). The problem of mental action. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing.* Frankfurt am Mainz: MIND Group.

Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of ”pure” consciousness. *Philosophy and the Mind Sciences*, 1(I), 7.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural networks*, *9*(8), 1265-1279.

Miller, M., & Clark, A. (2018). Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, *195*(6), 2559-2575.

Miller, M., Kiverstein J. & Rietveld K. (forthcoming). Embodying Addiction: A Predictive Processing Framework. *Brain and Cognition.*

Millidge, B. (2020). Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, *96*, 102348.

Millidge, B., Tschantz, A., & Buckley, C. L. (2021). Whence the expected free energy?. *Neural Computation*, 33(2), 447-482.

Millière, R. (2017). Looking for the self: Phenomenology, neurophysiology and philosophical significance of drug-induced ego dissolution. *Frontiers in Human Neuroscience*, 11, 245.

Millière, R. (2020). The varieties of selflessness. *Philosophy and the Mind Sciences*, *1*(I), 1–41.

Milliere, R., & Metzinger, T. (2020). Radical disruptions of self-consciousness. *Philosophy and the Mind Sciences*, *1*(I), 1–13.

Millière, R., Carhart-Harris, R. L., Roseman, L., Trautwein, F. M., & Berkovich-Ohana, A. (2018). Psychedelics, meditation, and self-consciousness. *Frontiers in psychology*, *9*, 1475.

Millikan, R. (1995). Pushmi-pullyu representations. *Philosophical Perspectives* 9:185-200.

Milner, D., & Goodale, M. (2006). *The visual brain in action* (Vol. 27). OUP Oxford

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Frontiers in computational neuroscience*, *10*, 56.

Mirza, M. B., Adams, R. A., Mathys, C., & Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE*, *13*(1), 1–20.

Montague, P. R., & King-Casas, B. (2007). Efficient statistics, common currencies and the problem of reward-harvesting. *Trends in Cognitive Sciences*, 11(12), 514–519.

Moreno, F. A., Wiegand, C. B., Taitano, E. K., & Delgado, P. L. (2006). Safety, tolerability, and efficacy of psilocybin in 9 patients with obsessive-compulsive disorder. Journal ofClinical Psychiatry, 67(11), 1735–1740.

Morsella, E. (2005). The function of phenomenal states: supramodular interaction theory. *Psychological Review*, *112*(4), 1000.

Morton, D. L., El-Deredy, W., Watson, A., & Jones, A. K. P. (2010). Placebo analgesia as a case of a cognitive style driven by prior expectation. *Brain Research*, *1359*(2005), 137–141.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, 25, 67–76.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, 66(3), 241–251.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological cybernetics*, *66*(3), 241-251.

Myowa-Yamakoshi, M., & Takeshita, H. (2006). Do human fetuses anticipate self- oriented actions? A study by four-dimensional (4D) ultrasonography. Infancy, 10(3), 289–301.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

Ñāṇamoli and Bodhi (1995). The middle length discourses of the Buddha. Boston: Wisdom.

Noe, A., & O'Regan, J. K. (2003). On comprehending the sensory effects of movement: Toward a theory of perception and consciousness. *Perception ECVP abstract*, *32*, 0-0.

Norcia, A. (2006) The Coffer Illusion.. In F. Macpherson (ed.), *The Illusions Index*. Retrieved from https://www.illusionsindex.org/i/coffer-illusion.

Nour, M. M., & Carhart-Harris, R. L. (2017). Psychedelics and the science of self-experience. *The British Journal of Psychiatry*, 210(3), 177–179.

Nour, M. M., Evans, L., Nutt, D., & Carhart-Harris, R. L. (2016). Ego-dissolution and psychedelics: Validation of the ego-dissolution inventory (EDI*). Frontiers in Human Neuroscience*, 10, 269.

Nutt, D. J., King, L. A., & Phillips, L. D. (2010). Drug harms in the UK: A multicriteria decision analysis. *The Lancet*, 376(9752), 1558–1565.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939-973.

O'Brien, G., & Opie, J. (2003). The multiplicity of consciousness and the emergence of the self. *The Self in Neuroscience and Psychiatry*, 107–120.

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams, R. B., & Bar, M. (2016). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition*, *47*, 63–74.

O'Regan, J. K., Myin, E., & Noë, A. (2003). Phenomenal consciousness explained (better) in terms of corporality and alerting capacity. *Consciousness & Cognition*.

O'Sullivan, N., de Bezenac, C., Piovesan, A., Cutler, H., Corcoran, R., Fenyvesi, D., & Bertamini, M. (2018). I am there… but not quite: an unfaithful mirror that reduces feelings of ownership and agency. *Perception*, *47*(2), 197–215.

Oades, R. D., & Isaacson, R. L. (1978). The development of food search behavior by rats: The effects of hippocampal damage and haloperidol. *Behavioral Biology*, *24*(3), 327–337.

Oakley, Justin, 1992. *Morality and the Emotions*, London: Routledge and Kegan Paul.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, *42*(3), 145-175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Ongaro, G., & Kaptchuk, T. J. (2019). Symptom perception, placebo effects, and the Bayesian brain. *Pain*, *160*(1), 1

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. Philosophy of Mind.

Orlandi, N. (2016). Bayesian perception is ecological perception. *philosophical topics*, *44*(2), 327-352.

Oudeyer, P. Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, *11*(2), 265-286.

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*(1), 27-37.

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. *Journal of Theoretical Biology*, *486*, 110089.

Palhano-Fontes, F., Barreto, D., Onias, H., Andrade, K. C., Novaes, M. M., Pessoa, J. A., et al. (2019). Rapid antidepressant effects of the psychedelic ayahuasca in treatment-resistant depression: A randomized placebo-controlled trial. *Psychological Medicine*, 49(4), 655–663.

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.

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

Panksepp, J. (1998). The periconscious substrates of consciousness: Affective states and the evolutionary origins of the self. *Journal of Consciousness Studies*, *5*(5–6), 566–582.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, *14*(1), 30–80.

Panksepp, J. (2008). The affective brain and core consciousness: how does neural activity generate emotional feelings?

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, *7*(1), 1–21.

Parr, T., & Friston, K. J. (2019). Generalised free energy and active inference. Biological cybernetics, 113(5), 495-513

Parr, T., Rees, G., & Friston, K. (2018). Computational neuropsychology and Bayesian inference. *Frontiers in Human Neuroscience*, *12*(February), 61.

Patočka, J. (1998). Body, Community, Language, World, ed. *J. Dodd. Chicago & La Salle, IL: Open Court.*

Pavlov, I. P. (1927). Conditional reflexes: an investigation of the physiological activity of the cerebral cortex.

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. Phenomenology and the Cognitive Sciences, 5(3-4), 229–269.

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(3), 902-911.

Pezzulo, G. (2017). Tracing the Roots of Cognition in Predictive Processing. *Open MIND*, 1–20.

Pezzulo, G. (2018). Commentary: The Problem of Mental Action: Predictive Control Without Sensory Sheets. *Frontiers in Psychology*, *9*, 1291.

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.

Pezzulo, G., & Nolfi, S. (2019). Making the Environment an Informative Place: A Conceptual Analysis of Epistemic Policies and Sensorimotor Coordination. *Entropy*, *21*(4), 350.

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L., & Friston, K. (2016). Active Inference, epistemic value, and vicarious trial and error. *Learning & Memory*, *23*(7), 322-338

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35

Pezzulo, G., Rigoli, F., & Friston, K. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, 22(4), 294–306.

Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT press.

Picard, F., & Friston, K. (2014). Predictions, perception, and a sense of self. Neurology, 83(12), 1112–1118.

Pink-Hashkes, S., Rooij, I. van, & Kwisthout, J. (2017). Perception is in the details: A predictive coding account of the psychedelic phenomenon. Proceedings of the 39th annual meeting ofthe cognitive science society, 2907–2912.

Poulet, J. F. A., & Hedwig, B. (2002). A corollary discharge maintains auditory sensitivity during sound production. *Nature*, *418*(6900), 872–876.

Powers, W. T., & Powers, W. T. (1973). *Behavior: The control of perception* (p. ix). Chicago: Aldine.

Preller, K. H., & Vollenweider, F. X. (2016). Phenomenology, structure, and dynamic of psychedelic states. In A. L. Halberstadt, F. X. Vollenweider, & D. E. Nichols (Eds.), *Behavioral neurobiology of psychedelic drugs* (pp. 221–256).

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford University Press.

Ramachandran, V., Chunharas, C., Marcus, Z., Furnish, T., & Lin, A. (2018). Relief from intractable phantom pain by combining psilocybin and mirror visual-feedback (MVF). *Neurocase*, 24(2), 105–110.

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225–239.

Ramstead, M. J. D., 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.

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

Richards, W. A. (2015). Sacred knowledge: Psychedelics and religious experiences. New York: Columbia University Press.

Ridderinkhof, K. R., Van Den Wildenberg, W. P., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and cognition*, *56*(2), 129-140.

Rochat, P. (1998). Self-perception and action in infancy. *Experimental Brain Research*, 123(1–2), 102–109.

Roepstorff, A. (2013). Interactively human: Sharing time, constructing materiality.

Romanes, G. J. (1888). *Mental evolution in man: Origin of human faculty*. Kegan Paul, Trench.

Romano, A. G., Quinn, J. L., Li, L., Dave, K. D., Schindler, E. A., Aloyo, V. J., & Harvey, J. A. (2010). Intrahippocampal LSD accelerates learning and desensitizes the 5-ht 2A receptor in the rabbit, romano et al. Psychopharmacology, 212(3), 441–448.

Roseman, L., Nutt, D. J., & Carhart-Harris, R. L. (2018). Quality of acute psychedelic experience predicts therapeutic efficacy of psilocybin for treatment-resistant depression. *Frontiers in Pharmacology*, 8, 974.

Rösler, L., Rolfs, M., Van der Stigchel, S., Neggers, S. F. W., Cahn, W., Kahn, R. S., & Thakkar, K. N. (2015). Failure to use corollary discharge to remap visual target locations is associated with psychotic symptom severity in schizophrenia. *Journal of Neurophysiology*, *114*(2), 1129–1136.

Rösler, L., Rolfs, M., Van der Stigchel, S., Neggers, S. F., Cahn, W., Kahn, R. S., & Thakkar, K. N. (2015). Failure to use corollary discharge to remap visual target locations is associated with psychotic symptom severity in schizophrenia. *Journal of Neurophysiology*, 114(2), Ruvolo, P., Messinger, D., & Movellan, J. (2015). Infants time their smiles to make their moms smile. PloS One, 10(9).

Safron, A. (2020). An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic . *Frontiers in Artificial Intelligence*, *3*(June).

Sakai, K. L. (2005). Language acquisition and brain development. *Science*, *310*(5749), 815-819

Salomon, R., Lim, M., Herbelin, B., Hesselmann, G., & Blanke, O. (2013). Posing for awareness: Proprioception modulates access to visual consciousness in a continuous flash suppression task. *Journal of Vision*, *13*(7), 2.

Salomon, R., Ronchi, R., Dönz, J., Bello-Ruiz, J., Herbelin, B., Martet, R., … Blanke, O. (2016). The Insula Mediates Access to Awareness of Visual Stimuli Presented Synchronously to the Heartbeat. *The Journal of Neuroscience*, *36*(18), 5115–5127.

Sanfey, A. G., Loewenstein, G., McClure, S. M., and Cohen, J. D. (2006). Neuroeconomics: cross-currents in research on decision-making. *Trends Cogn. Sci.* 10, 108–116.

Sass, L. A., & Parnas, J. (2003). Schizophrenia, consciousness, and the self. Schizophrenia bulletin, 29(3), 427-444.

Savage, C. (1955). Variations in ego feeling induced by D-lysergic acid diethylamide (LSD-25). Psychoanalytic Review, 42(1), 1–16.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379-399.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic simulation of future events: Concepts, data, and applications. Annals of the New York Academy of Sciences, 1124(1), 39–60.

Schillaci, G., Ritter, C. N., Hafner, V. V., & Lara, B. (2016, July). Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents. In *Proceedings of the Artificial Life Conference 2016 13* (pp. 390-397). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230-247.

Schomaker, L. (2004, October). Anticipation in cybernetic systems: A case against mindless anti-representationalism. In *2004 IEEE International Conference on Systems, Man and Cybernetics*

*(IEEE Cat. No. 04CH37583)* (Vol. 2, pp. 2037-2045). IEEE.

Schulkin, J., & Sterling, P. (2019). Allostasis: A brain-centered, predictive mode of physiological regulation. *Trends in Neurosciences*, 42(10), 740–752.

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in psychology*, *4*, 710.

Scott, G., & Carhart-Harris, R. L. (2019). Psychedelics as a treatment for disorders of consciousness. Neuroscience of Consciousness, 2019(1), niz003.

Sebastián, M. Á. (2020). Perspectival self-consciousness and ego-dissolution: An analysis of (some) altered states of consciousness. *Philosophy and the Mind Sciences*, 1(I), 9.

Seeley, T. D. (2009). *The wisdom of the hive: the social physiology of honey bee colonies*. Harvard University Press.

Segal, S. (2002). Collision with the infinite: A life beyond the personal self. New Age Books.

Seth, A. (2014). The cybernetic bayesian brain - from interoceptive inference to sensorimotor contingencies. *Open MIND*, *35*, 1–24.

Seth, A. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. Windt (Eds.), Open MIND (pp. 1–24). Frankfurt am Main: MIND Group.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565-573

Seth, A. K. (2014). *The cybernetic Bayesian brain*. Open MIND. Frankfurt am Main: MIND Group

Seth, A. K. (2018). Consciousness: The last 50 years (and the next). *Brain and Neuroscience Advances*, *2*, 239821281881601.

Seth, A. K., & Friston, K. (2016). Active interoceptive inference and the emotional brain. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1708), 20160007.

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

Seth, A. K., & Hohwy, J. (2020). Predictive processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, 1-2.

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: the somatic basis of selfhood. *Trends in cognitive sciences*, *22*(11), 969-981.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, *2*, 395.

Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience*, 33, 89–108.

Sharp, P. E. (2014). Meditation-induced bliss viewed as release from conditioned neural (thought) patterns that block reward signals in the brain pleasure center. *Religion, Brain & Behavior*, *4*(3), 202-229.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, 7, 1792.

Shipp, S., Adams, R., & Friston, K. J. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences*, *36*(12), 706–16

Shulman, E. (2014). *Rethinking the buddha: early buddhist philosophy as meditative perception*. Cambridge University Press.

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., et al. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. Psychological Bulletin, 144(4), 343.

Simeon, D., & Abugel, J. (2006). *Feeling unreal: depersonalization disorder and the loss of the self.* Oxford University Press, USA.

Sirigu, A., Daprati, E., Pradat-Diehl, P., Franck, N., & Jeannerod, M. (1999). Perception of self-generated movement following left parietal lesion. *Brain*, *122*(10), 1867–1874.

Smith, J. (2017). Self-consciousness.

Smith, L. S., Hesp, C., Lutz, A., Mattout, J., Friston, K., & Ramstead, M. (2020). Towards a formal neurophenomenology of metacognition: modelling meta-awareness, mental action, and attentional control with deep active inference.

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019). Neurocomputational mechanisms underlying emotional awareness: insights afforded by deep active inference and their potential clinical relevance.

Solms, M. (2019). The Hard Problem of Consciousness and the Free Energy Principle, *9*(January), 1–16.

Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. WW Norton & Company.

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

Srofe, L. A., & Waters, E. (1976). The ontogenesis of smiling and laughter: a perspective on the organization of development in infancy. *Psychological review*, 83(3), 173.

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

Stephan, K. E., Manjaly, Z. M., Mathys, C., Weber, L. A., Paliwal, S., Gard, T., et al. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10, 550.

Stephens, G. L., & Graham, G. (1994). Self-consciousness, mental agency, and the clinical psychopathology of thought insertion. *Philosophy, Psychiatry, & Psychology*, *1*(1), 1–10.

Sterling P. Laughlin S. (2015). *Principles of Neural Design* . Cambridge: MIT Press

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

Sterling, P. (2012). Physiology & Behavior Allostasis : A model of predictive regulation. *Physiology & Behavior*, *106*(1), 5–15.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., et al. (2018). The predictive coding account of psychosis. Biological Psychiatry, 84(9), 634–643.

Strassman, R. J. (1984). Adverse reactions to psychedelic drugs. A review of the literature. Journal ofNervous and Mental Disease, 172(10), 577–595.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in cognitive sciences*, *13*(9), 403-409.

Thakkar, K. N., Mathalon, D. H., & Ford, J. M. (2021). Reconciling competing mechanisms posited to underlie auditory verbal hallucinations. *Philosophical Transactions of the Royal Society B*, *376*(1817), 20190702.

Thompson E (2007) *Mind in life: biology, phenomenology, and the sciences of mind.* Harvard University Press, Cambridge, MA

Timmermann, C., Roseman, L., Schartner, M., Millière, R., Williams, L., Erritzoe, D., et al. (2019). Neural correlates of the DMT experience as assessed via multivariate EEG. bioRxiv, 706283.

Timmermann, C., Spriggs, M. J., Kaelen, M., Leech, R., Nutt, D. J., Moran, R. J., et al. (2018). LSD modulates effective connectivity and neural adaptation mechanisms in an auditory oddball paradigm. Neuropharmacology, 142, 251–262.

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLOS Computational Biology*, *16*(4), e1007805.

TURING, I. B. Y. A. M. (1950). Computing machinery and intelligence-AM Turing. Mind, 59(236), 433.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649-665.

Van de Cruys, S. (2017). Affective value in the predictive mind. MIND Group.

Van de Cruys, S., & Wagemans, J. (2011). Putting reward in art: a tentative prediction error account of visual art. i-Perception, 2(9), 1035-1062.

Varela, F. J. (1991). Organism: A meshwork of selfless selves. *In Organism and the Origins of Self (pp. 79-107). Springer, Dordrecht.*

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience*. MIT press.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking Through Other Minds: A Variational Approach to Cognition and Culture. *Behavioral and Brain Sciences*, (May).

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.

Walsh, K.S., McGovern, D.P., Clark, A. and O'Connell, R.G. (2020), Evaluating the neurophysiological evidence for predictive processing as a model of perception. Ann. N.Y. Acad. Sci..

Ward, D., Roberts, T., & Clark, A. (2011). Knowing what we can do: actions, intentions, and the construction of phenomenal experience. *Synthese*, *181*(3), 375–394.

Wehner, R. (2013). Life as a cataglyphologist—and beyond. *Annual Review of Entomology*, *58*, 1–18.

Whitwell, R. L., Milner, A. D., & Goodale, M. A. (2014). The two visual systems hypothesis: new challenges and insights from visual form agnosic patient DF. *Frontiers in Neurology*, *5*, 255.

Whybrow, P. C. (2015). *The Well-Tuned Brain: The Remedy for a Manic Society*. WW Norton & Company.

Whyte, C. J. (2019). Integrating the global neuronal workspace into the framework of predictive processing: Towards a working hypothesis. *Consciousness and Cognition*, *73*(April), 102763.

Whyte, C. J., & Smith, R. (2020). The predictive global neuronal workspace: A formal active inference model of visual consciousness. *Progress in Neurobiology*, 101918.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. Technology Press.

Wiese, W. (2014). Perceptual presence in the Kuhnian-Popperian Bayesian brain. In *Open Mind.* Open MIND. Frankfurt am Main: MIND Group.

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

Wiese, W. (2018). Experienced Wholeness, (January 2019), 293–294.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*.

Wilkinson, S., Deane, G., Nave, K., & Clark, A. (2019). Getting warmer: predictive processing and the nature of emotion. In *The Value of Emotions for Knowledge* (pp. 101-119). Palgrave Macmillan, Cham.

Williford, K., Bennequin, D., Friston, K., & Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Frontiers in Psychology*, *9*(DEC), 1–18.

Winter, U., LeVan, P., Borghardt, T. L., Akin, B., Wittmann, M., Leyens, Y., & Schmidt, S. (2020). Content-free awareness: EEG-fcMRI correlates of consciousness as such in an expert meditator. *Frontiers in Psychology*, *10*, 3064.

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current biology*, *11*(18), R729-R732

Yamada, Y., Kanazawa, H., Iwasaki, S., Tsukahara, Y., Iwata, O., Yamada, S., & Kuniyoshi, Y. (2016). An embodied brain model of the human foetus. Scientific Reports, 6, 27893

Yoshida, W., Dolan, R. J., & Friston, K. (2008). Game theory of mind. PLoS *Computational Biology*, 4(12), e1000254.

Zahavi, D. (2014). *Self and other: Exploring subjectivity, empathy, and shame*. Oxford University Press, USA.

Zhou, W., Jiang, Y., He, S., & Chen, D. (2010). Olfaction modulates visual perception in binocular rivalry. *Current Biology*, *20*(15), 1356–1358.

Zoia, S., Blason, L., D'Ottavio, G., Bulgheroni, M., Pezzetta, E., Scabar, A., & Castiello, U. (2007). Evidence of early development of action planning in the human foetus: a kinematic study. *Experimental Brain Research*, 176(2), 217–226.