Predictive Coding and Neurocomputational Psychiatry: A Mechanistic Framework for Understanding Mental Disorders

Alexander D Shaw[1], Rachael L Sumner[2], and Lioba C S Berndt[1]

[1]Dept. of Psychology, Faculty of Health & Life Sciences, University of Exeter, UK
[2]School of Pharmacy, University of Auckland, NZ

Abstract

Predictive coding offers a powerful computational framework for understanding brain function and psychiatric disorders at a mechanistic level. This perspective synthesizes advances in computational psychiatry, proposing that mental disorders can be conceptualized as specific alterations in the brain's predictive inference machinery. We first outline the theoretical foundations of predictive coding, including Bayesian inference, free-energy minimization, and neural population dynamics, showing how these abstract computational principles map onto specific neural circuits and biophysical mechanisms. We then demonstrate how diverse psychiatric conditions can be understood within this unified framework. Critically, this additionally provides a basis upon which predictive coding becomes a testable, modifiable, falsifiable construct within biological psychiatry.

Beyond offering conceptual clarity, this framework has significant clinical implications, including the development of mechanistic biomarkers, personalized treatment approaches based on computational phenotypes, and novel therapeutic interventions targeting specific inferential abnormalities. By grounding psychiatric symptoms in aberrant predictive processes implemented in neural circuitry, this approach promises a more mechanistic understanding of mental disorders and a path toward more targeted, effective interventions.

1 Introduction

Traditional psychiatry has long relied on symptom-based classifications of mental illness, with limited insight into underlying brain mechanisms. Computational psychiatry offers a new paradigm [1, 2, 3, 4]: by integrating mathematical models with neurobiology, it seeks to explain psychiatric phenomena in terms of aberrant brain computations [5]. Among various modelling approaches, predictive coding has emerged as a compelling unifying theory for brain function [6, 7]. In essence, predictive coding treats the brain as an inference machine that continually generates and updates predictions about sensory inputs. This framework provides a normative (Bayesian) account of perception and action, where the brain tries to minimise surprise or "free energy" by aligning its internal model with the outside world [8]. Crucially, this perspective offers a plausible account of neuronal computation in cortical circuits: hierarchical networks of neurons are thought to exchange top-down predictions and bottom-up prediction errors to achieve efficient information processing.

At the same time, predictive coding provides a principled way to think about mental illness. Many psychiatric symptoms can be interpreted as errors of inference - that is, failures in the predictive coding machinery [9]. For example, hallucinations and delusions in psychosis may result from placing too much weight on prior beliefs (or not enough on sensory evidence) - a computationally precise deficit that could serve as a quantitative biomarker for psychosis subtypes. Conversely, autistic perception may stem from overly weak prior expectations leading to sensory overload, representing a fundamentally different computational phenotype. This review will examine how major psychiatric conditions can be framed in terms of aberrant predictive coding, linking computational deviations to clinical phenomena.

Finally, predictive coding is especially attractive because it maps onto neural circuits. The abstract variables of predictive models (predictions, prediction errors, precision weights) can be associated with specific neuron populations and connections in cortical microcircuits. These concrete mappings mean that predictive coding hypotheses are testable with neurobiological data [8]. Using neural mass models and Dynamic Causal Modelling (DCM), researchers can design

generative models of brain activity and compare them to recordings (M/EEG, fMRI) to quantify these computational parameters in individual patients, moving toward personalized computational phenotyping in psychiatry. Since DCM was first introduced in 2003 [10], the models themselves are being refined, expanded and tested.

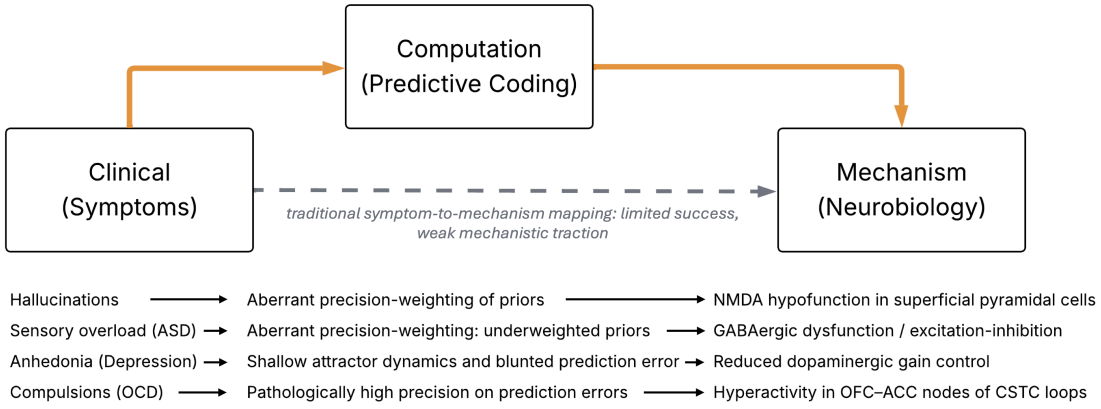

Figure 1: Predictive coding as a computational bridge between psychiatric symptoms and neurobiological mechanisms. Traditional approaches (grey dashed line) attempt to map clinical phenomena directly onto biological substrates, often yielding limited mechanistic insight. Predictive coding provides a principled intermediate layer (orange arrows), allowing symptoms to be reframed as computational inference failures-such as aberrant precision-weighting or dysfunctional attractor dynamics-which can in turn be grounded in specific circuit-level or receptor-level pathophysiology. Examples illustrate this tri-level mapping for hallucinations, sensory overload, anhedonia, and compulsions.

In what follows, we outline the current state of the art of the predictive coding framework and its neural implementations, then explore how it illuminates psychiatric disorders, and finally discuss the broader implications for diagnosis and treatment in psychiatry.

2 Theoretical Foundations of Predictive Coding in the Brain

2.1 Brain as an Inference Machine – Bayesian Reasoning and Dynamics:

A central premise of predictive coding (and related theories like the Bayesian brain hypothesis) is that the brain performs some form of Bayesian inference [11]. The brain maintains internal generative models that produce predictions about sensory inputs, and it updates these models when predictions fail [12, 7]. Mathematically, we can describe the brain's state dynamics and outputs with differential equations [10]. For example, a simple state-space model of neural activity can be written as:

$$\frac{dx}{dt} = f(x,\, u,\, P), \quad y = g(x,\, P), \tag{1}$$

where x represents internal neural states (variables), u external inputs, P model parameters, and y observable outputs (e.g. neural signals). Here, f encapsulates how neural states evolve over time (governed by physiology), and g maps internal states to observed signals.

In healthy function, these dynamics settle into stable regimes (attractors) that support adaptive perceptions and behaviours. In contrast, psychiatric disorders may correspond to pathological attractor states in this system – for instance, depression might involve a neural state getting "stuck" in a low-firing, inflexible regime. By formulating such models and fitting them to patient data (e.g. M/EEG), we can infer which physiological parameters (P) are altered in illness (for example, synaptic gain or connectivity) and thereby gain mechanistic insight [10].

2.2 Predictive Coding and Free-Energy Minimisation

Predictive coding extends the above by positing a specific computational strategy for the brain: minimise the error between expected and actual inputs [13, 14]. In a predictive coding scheme, higher brain regions send predictions (top-down signals) about lower-level activity, and lower

regions compute prediction errors (differences between what was predicted and what is actually sensed) to send back upward. The brain then adjusts its internal states to reduce these errors. A simple formulation of the predictive coding update is:

$$\Delta\hat{\mu} = \eta\epsilon, \tag{2}$$

where $\hat{\mu}$ is the brain's current prediction (or estimate of a latent cause), ϵ is the prediction error (difference between observed input y and the predicted input $\hat{y}$), and η is a learning rate.

In other words, the estimate is adjusted in proportion to the error signal. This can be seen as a gradient ascent step on an implicit log-likelihood or a descent on "surprise." By iteratively refining its predictions in this manner, the brain approaches a state that maximises model evidence and minimises surprise (or variational free energy). Notably, this framework is formally equivalent to Bayesian inference: the brain's updated estimate $\hat{\mu}$ comes to approximate the posterior belief that combines prior expectation with (sensory) likelihood. The Free Energy Principle generalises this idea, proposing that neural dynamics minimise a free-energy bound on the discrepancy between the brain's model and sensory data [12]. This theoretical principle bridges computation and neurobiology: under certain assumptions (e.g. Gaussian noise), it yields biologically plausible rules for synaptic updates that implement Bayesian belief updating in neural circuits.

2.3 Neural Populations and Mean-Field Models:

Brain networks consist of large populations of neurons making detailed single-cell modeling computationally prohibitive. To connect microscopic neural activity with macroscopic brain signals (e.g. M/EEG or fMRI), predictive coding models often invoke a mean-field or neural mass approximation [15]. Rather than track every neuron, one tracks the average activity $\langle x \rangle$ of a population, along with summary statistics of variability. For example, a mean-field equation might describe the evolution of the population firing rate, while a second equation captures the variance or correlations within the population. Such population models are integral to implementations of predictive coding in the cortex: they allow one to treat an entire cortical column or region as a unit that sends and receives prediction/error signals [16]. Importantly, the hierarchical organisation – populations organised in layers and areas – naturally lends itself to the hierarchical Bayesian structure of predictive coding [17]. Each level of the hierarchy deals with a different scale of representation, and population dynamics at that level encode predictions or errors about that content. By adjusting a few key parameters (like the gain of neuronal populations that carry error signals), these models can simulate how precision (confidence) is encoded and modulated in the brain [18, 19]. This provides a way to link neurotransmitter systems (e.g. NMDA, dopamine or serotonin, which affect synaptic gain) to computational quantities in predictive coding (like precision-weighting of prediction errors). Such connections are central to understanding psychiatric conditions in this framework where altered neuromodulatory systems may directly impact precision-weighting and inference.

2.4 Biophysical Generative Models – Hodgkin–Huxley and Beyond:

An attractive aspect of predictive coding theory is that it can be grounded in detailed neurobiology. The function $f(x)$ in our state equations (Eq.1) can be specified using well-established biophysical models of neurons. A classic example is the conductance-based Hodgkin–Huxley model [20], which describes how a neuron's membrane voltage V evolves over time due to ionic currents. In a Hodgkin–Huxley formulation:

$$C_m \frac{dV}{dt} = g_{\mathrm{Na}} m^3 h (E_{\mathrm{Na}} - V) + g_{\mathrm{K}} n^4 (E_{\mathrm{K}} - V) + g_L (E_L - V), \tag{3}$$

where C_m is the membrane capacitance, $g_{\mathrm{Na}}, g_{\mathrm{K}}, g_L$ are the maximal conductances of sodium, potassium, and leak channels, $E_{\mathrm{Na}}, E_{\mathrm{K}}, E_L$ are their reversal (Nernst) potentials, and m, h, n are gating variables that evolve according to their own dynamics. This detailed equation is a concrete instantiation of the function $f(x)$ governing neuronal dynamics, and can be easily extended to include a range of neurotransmitter systems (for example, adding calcium channels and a voltage gate to model NMDA; [21]).

While such biophysical complexity is often simplified in higher-level models, it reminds us that any computational theory like predictive coding ultimately must respect the laws of neurophysiology. To bridge single-neuron dynamics with neural population behavior, these biophysical principles can be incorporated into mean-field models, creating conductance-based neural masses

that maintain biological realism while allowing scalability. Indeed, one can build generative models that incorporate known physiology (e.g. receptor kinetics, membrane time constants) and then invert those models to explain observed neural data. Computational psychiatry studies have done exactly this – for instance, using conductance-based neural mass models to infer synaptic changes in disorders like schizophrenia from M/EEG recordings [22, 23].

3 Neuronal Circuitry of Predictive Coding

Building on the theoretical foundations discussed above, we now consider how predictive coding might be implemented in actual brain circuitry. A growing body of work suggests that canonical cortical circuits – the repeating layered networks in cortex – are well-suited to implement the hypothesised message-passing of predictions and errors. In a hierarchical predictive coding model, every cortical area (or layer) has units that encode the current prediction of some features and units that encode the prediction error (the unexplained residual). Physiologically, a plausible mapping is that deep-layer pyramidal neurons carry top-down predictions to lower areas, while superficial-layer pyramidal neurons carry forward prediction errors to higher areas (figure 2).

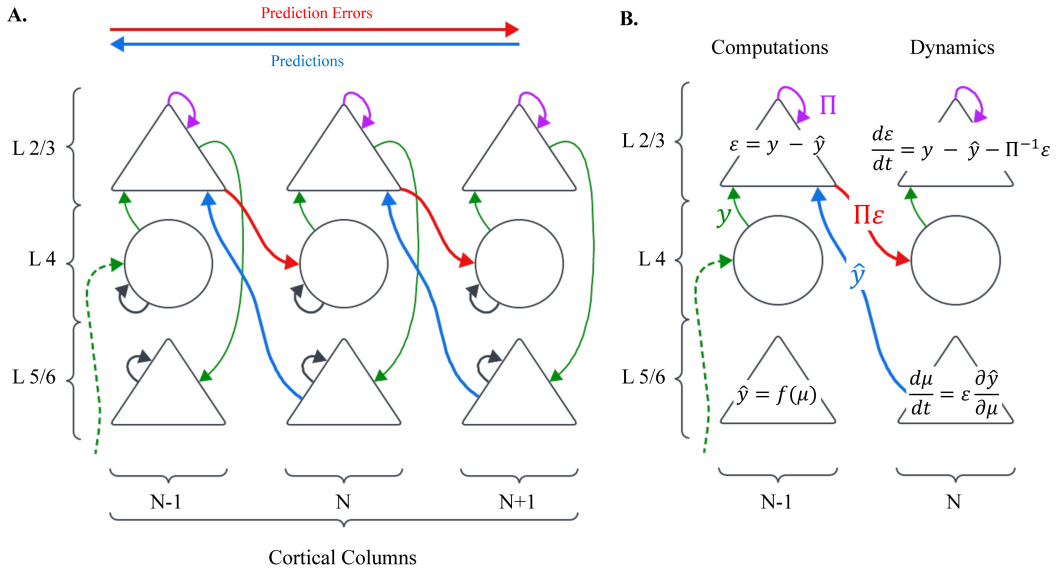

Figure 2: Schematic of hierarchical predictive coding across cortical levels. **A.** Higher areas (N+1) send predicted signals (feedback, blue arrows) to lower sensory areas (N), forming empirical priors for incoming data. At the lower level, the difference between actual input (green arrow) and the top-down prediction constitutes a prediction error. This error is weighted (purple arrow) and passed forward (feedforward) to update the higher-level representation (red arrow) in the form of a 'precision weighted prediction error'. **B.** The encoding of predictions and prediction errors plausibly takes place in deep (L5/6) and superficial (L2/3) pyramidal populations, respectively. Furthermore, since this process is dynamical (rather than fixed) these computations can be formulated as differential equations (far right).

Supporting this model, neuroanatomical studies have found a remarkable correspondence between the connectivity of cortical microcircuits and the connections implied by predictive coding theories [17]. Specifically, the laminar patterns of feedforward vs. feedback projections align with the idea that separate neuronal populations send predictions downward and errors upward in the hierarchy. For example, feedforward projections originate mainly from superficial layers (L2/3) and target middle-layer (L4) neurons in the next cortical area, whereas feedback projections originate from deep layers (L5/6) and target superficial layer neurons in the preceding area [24]. This matches the predictive coding requirement for distinct streams of information flow. A detailed model by Bastos and colleagues [17] formalised this, assigning biophysical neuron models to the roles of error units and prediction units and demonstrating consistency with observed cortical beta and gamma rhythms. Inhibitory interneurons also play crucial roles in this framework, potentially controlling precision-weighting through their modulatory effects on pyramidal

cell activity. Moreover, the involvement of the thalamus is often interpreted in predictive coding terms: the thalamus might help compare cortical predictions with incoming signals, or gate the precision of ascending sensory data [25]. Dysfunctions in thalamo-cortical loops are indeed implicated in disorders like schizophrenia and OCD, consistent with predictive coding abnormalities (e.g. thalamic filter failure leading to sensory overload or intrusive errors) [2].

Because *predictive coding neatly maps onto specific circuits*, we can use data-driven modelling to test hypotheses about those circuits. DCM is a prominent approach that evaluates network models against measured brain signals [26]. In DCM, one posits a circuit model (with directed connections, layers, etc.), and uses Bayesian inference to estimate the connection strengths that best explain the data for different groups or conditions. Predictive coding provides guidance for constructing such models (e.g. which connections should change under certain tasks or pathologies). A striking example comes from a DCM study on people with schizophrenia performing a perception task [5]. The modelling results showed that compared to healthy subjects, schizophrenia patients had markedly reduced backward connectivity from frontal cortex to visual cortex, and moreover, unlike controls, they failed to increase this top-down connectivity when stimuli became predictable. In other words, the normal adaptive tuning of cortical feedback based on predictability was absent in schizophrenia – exactly what one would expect if the brain's predictive coding machinery (which relies on adjusting top-down signals) was impaired. This illustrates how computational models and neural data can converge: by examining circuit parameters estimated via DCM, we obtain evidence that "precision-weighting of prediction errors is deficient" or "top-down predictions are underutilised" in a given disorder. 2336

4 Predictive Coding and Psychiatric Disorders

Using the predictive coding lens, we can reinterpret several major psychiatric conditions as specific forms of inference gone awry [27]. In each case, symptoms are linked to particular disturbances in how predictions, prediction errors, or their precision weights are handled in the brain [28]. Crucially, because of the mapping between predictive coding and cortical wiring, proposed changes in predictive coding represent testable hypotheses of pathophysiology with biologically meaningful parameters [9, 29]. Below, we review key examples and supporting evidence. 2419

4.1 Schizophrenia: Aberrant Precision-Weighting of Priors

Schizophrenia has been hypothesised to result from an imbalance in the brain's handling of prediction errors and priors. In healthy perception, the brain assigns an optimal precision (or confidence) to sensory evidence relative to prior beliefs, so that neither hallucinations (overly strong priors) nor confusion (overly strong sensory noise) occurs. In schizophrenia, this balance appears to be disrupted: individuals may place too much weight on internally generated predictions (priors) and not enough on external sensory input [2]. In predictive coding terms, there is an aberrant precision-weighting such that top-down signals are afforded inappropriately high precision relative to bottom-up signals. Formally, one can think of the prediction error term:

$$\epsilon = (y - \hat{y}) \tag{4}$$

being under-weighted, or conversely the prior's influence being over-weighted, due to mis-tuned precision (Π) on error neurons.

This idea helps explain classic positive symptoms: delusions can be seen as unfounded beliefs that persist because contradictory sensory evidence (prediction errors) is not given enough weight to overturn them. Hallucinations, likewise, could result from internally generated representations (predictions from higher cortex) intruding on perception because the brain is overly biased toward expecting its own hypothesis rather than the actual input [30]. In computational simulations, reducing the "precision of priors relative to sensory evidence" indeed produces hallucination-like phenomena.

At the neurobiological level, this pathology of precision-weighting has been linked to dysregulation in both glutamatergic and GABAergic systems. NMDA receptor dysfunction, which is implicated in schizophrenia, affects synaptic gain and thus the encoding of prediction error signals [23]. Complementing this, GABAergic abnormalities-evidenced by decreased occipital GABA concentrations in patients-further disrupt the excitation-inhibition balance crucial for appropriate precision-weighting [31]. These neurochemical imbalances manifest in circuit-level changes: individuals exhibit reduced visually induced gamma oscillation frequencies and impaired orientation discrimination, both linked to excitation–inhibition imbalances in visual cortex.

Empirical studies using tasks like oddball detection provide further evidence for this framework. Individuals with schizophrenia often show reduced mismatch negativity responses, consistent with improper error signalling [32]. DCM analyses reveal the circuit-level consequences of these abnormalities: individuals show diminished local synaptic connectivity, particularly between inhibitory interneurons and superficial pyramidal cells (the putative error units in predictive coding), with connectivity deficits correlating with negative symptom severity [31]. Furthermore, DCM analyses have found weakened feedback connectivity in cortical hierarchies, supporting the notion of a breakdown in top-down predictive stability. In summary, schizophrenia can be cast as a disorder of belief updating: the filters that should correct false beliefs via error signals are themselves corrupted, leaving aberrant beliefs (paranoia, hallucinations) unchecked [33]. While negative symptoms are not discussed in detail here, many (such as apathy or social withdrawal) may reflect dynamics that align more closely with depressive phenotypes-such as impaired flexibility or reduced prediction error sensitivity.

4.2 Autism: Weak Priors and Sensory Overweighting

Autism Spectrum Disorder (ASD) offers a contrasting case to schizophrenia within predictive coding accounts of cognition [34]. Although autism is not classified as a psychiatric disorder, it is frequently seen in neurodevelopmental and mental health services, and computational models of inference have been used to understand autistic perception and cognition. The influential theory by Pellicano and Burr [35] posits that autistic individuals form unusually weak priors, and this has been extended into predictive coding accounts that highlight atypical inference processes in ASD [36]. In Bayesian terms, the prior in autistic perception is assigned low precision, leading to perceptual hypersensitivity and a focus on raw input [37]. In other words, the posterior estimate is driven primarily by the likelihood (sensory evidence), with minimal top-down constraint.

Formally, we can express the imbalance as a breakdown in Bayes' rule: the posterior $\sim$ likelihood $\times$ prior, but if $P(\text{prior})$ is assigned a very low weight, the brain relies almost entirely on incoming data. One simple equation reflecting this might be:

$$Posterior \propto P(\text{observation}|\text{state}) \times [P(\text{state})(very\ small)] \tag{5}$$

Behaviourally, this manifests as an insistence on *sameness* and difficulty generalising from past experience (since each situation is processed afresh, without strongly applying past lessons). It also aligns with enhanced local processing versus impaired global integration (the autistic brain doesn't use broad priors to "fill in" gaps) [38, 39].

Neurophysiologically, researchers have noted an excitation/inhibition imbalance in ASD cortical circuits [40, 41]. One interpretation is that inhibitory processes that help implement predictions (by suppressing predictable inputs) are weaker, leading to a surfeit of unsuppressed (unexpected) signal. This may explain why many autistic individuals experience sensory overwhelm in environments with high levels of stimulation – their brains are not effectively filtering out predictable background information. Additionally, predictive coding accounts of autism highlight differences in precision modulation by neuromodulators like serotonin or acetylcholine, which may underlie the diminished influence of priors [42]. Interestingly, the proposed circuit abnormalities in autism – involving reduced influence of feedback connections and altered gain control in superficial pyramidal cells – mirror those in schizophrenia but with different parameter settings, suggesting these conditions may represent opposing ends of a predictive coding spectrum [37]. While the full picture is complex, the overarching view is that autism involves atypical inference-perception and learning that lean too heavily on the immediate evidence and struggle to incorporate the abstract, probabilistic regularities that typically guide human perception.

4.3 Depression: Maladaptive Attractor States and Failure to Update

Major depression is often characterised by rigid negative beliefs and an inability to adapt to positive new information. In predictive coding terms, one can think of depression as the brain getting trapped in a maladaptive predictive model that it fails to update despite contrary evidence. From a dynamical systems perspective, this corresponds to a pathological attractor state in neural activity [43, 44]. The brain's state-space dynamics settle into a basin of low energy (a pessimistic, self-consistent model of the world) and exhibit reduced flexibility to escape that basin. Formally, we might write the neural state evolution as:

$$\frac{dx}{dt} = f(x) \tag{6}$$

and say that $f(x)$ has an attractor that represents a depressive state. Normally, prediction errors (surprising positive events, changes in environment) would perturb the system out of that attractor, leading to an update of beliefs (e.g. learning that things aren't as hopeless as expected). In depression, however, there seems to be a failure to effectively minimise free energy with respect to new inputs. The brain's model does not get updated by unexpected good news; instead, the system may downweight those prediction errors or interpret them in a way that conforms to the negative prior belief [45] (e.g. "it was a fluke," "it won't last").

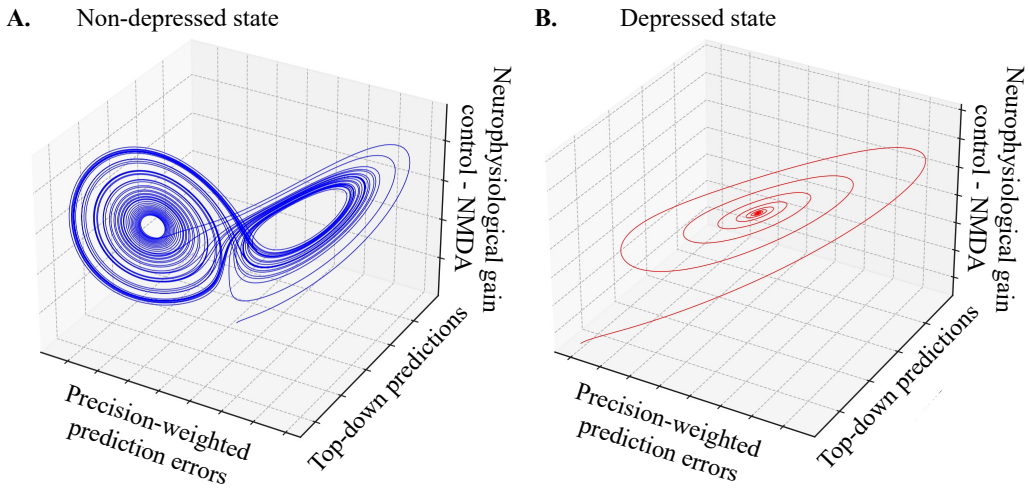

Figure 3: This figure illustrates differences in neural dynamics between healthy (**A**) and depressed (**B**) states using a toy attractor framework (The Lorenz attractor). The three axes represent key computational variables in predictive coding: **X-axis:** Precision-weighted prediction error (how much sensory input updates beliefs). **Y-axis:** Top-down priors (strength of internal expectations, including cognitive biases). **Z-axis:** Neuromodulatory gain control (influence of neurotransmitters like NMDA on precision). In the healthy state (blue attractor), neural dynamics are chaotic and flexible, allowing for continuous updating of beliefs in response to sensory input. The system explores a broad range of states, reflecting an adaptive balance between prior beliefs and prediction errors. In contrast, the depressed state (red attractor) exhibits a collapsed, deep attractor basin, indicative of rigid, maladaptive inference. Here, overweighted priors (high Y) dominate, while prediction error updates are suppressed (low X), leading to inflexible belief updating. Additionally, reduced neuromodulatory gain (low Z) diminishes the system's ability to dynamically adjust sensory precision, mirroring reduced serotonergic and dopaminergic function in depression.

Empirical evidence supports disrupted predictive processing in depression. For instance, depressed individuals show blunted neural responses to unexpected rewards (a kind of reduced reward prediction error signalling) [46]. This is consistent with anhedonic depression, where people no longer feel interest or reward in things they had previously. On the circuit level, depression has been associated with impaired top-down connectivity and reduced neuromodulatory drive (e.g. serotonin), which could both contribute to a sluggish predictive coding system that does not revise its priors promptly [47].

Unlike the positive symptoms of schizophrenia (with its overweighted priors) or autism (with its underweighted priors), depression may represent a different type of predictive coding dysfunction-one characterized by normal weighting but impaired updating dynamics. In the free-energy framework, one could say the depressive brain is not adequately exploring the space of alternative hypotheses; it is stuck with a high *model evidence* for a dark outlook because it isn't sampling the environment in an unbiased way. This view also resonates with psychological observations like confirmation bias in depression (selectively attending to negative feedback). At the cognitive level, the negative schema described in cognitive therapy can be recast as a prior with excessive precision that resists updating. This perspective suggests specific therapeutic approaches. Treatments like behavioral activation, which deliberately exposes individuals to potentially rewarding activities

[48], could work by forcing the sampling of environments that generate prediction errors contrary to the depressive model. Similarly, cognitive therapy may function by explicitly challenging the high-precision priors that maintain the depressive attractor state [49].

4.4 Obsessive–Compulsive Disorder (OCD): Overactive Error Signals and Precision

Obsessive–compulsive disorder can be understood as an exaggeration of the brain's error signals and uncertainty responses. In OCD, patients are tortured by a feeling that *something is wrong* or incomplete, leading to repetitive behaviours to mitigate that anxiety. Computationally, this maps to excessive precision-weighting of prediction errors, especially in the realm of threat or danger predictions [50, 51].

$$\frac{d\mu}{dt} = \varepsilon \cdot \frac{\partial \hat{y}}{\partial \mu} \tag{7}$$

This equation describes how internal beliefs (μ) are updated based on prediction errors ($\varepsilon = y - \hat{y}$) and the sensitivity of predicted input $\hat{y}$ to those beliefs. In OCD, the precision assigned to ε becomes pathologically high, effectively amplifying small discrepancies and driving excessive belief updates, even when the external world offers reassurance [52].

Even when things are objectively fine, the OCD brain generates a strong error signal indicating a discrepancy (e.g. "maybe my hands are still not clean" or "the door isn't truly locked"). These error signals are given abnormally high confidence, compelling the individual to act on them (wash again, check again). Essentially, the system has too low a threshold for declaring a prediction error and then cannot easily cancel that error signal.

One way to formalise this is to say that the complexity term in model optimisation dominates: the brain's model remains extremely complex and unwilling to accept that a simpler explanation (e.g. "the stove is off and nothing bad will happen") is sufficient. The hierarchy in predictive coding might get "stuck" with intermediate-level error units firing constantly. This distinguishes OCD from other disorders we've discussed: OCD involves appropriate weighting but with excessive precision on specific categories of errors, particularly those related to safety, contamination, or moral concerns.

Neurobiologically, OCD has been strongly linked to hyperactivity in specific circuits like the cortico-striatal-thalamo-cortical (CSTC) loop, particularly involving the orbitofrontal cortex (OFC) and anterior cingulate cortex (ACC) [53, 54]. These regions are associated with monitoring for errors and adjusting behaviour. Imaging studies show that at baseline and during symptom provocation, OFC and ACC are over-active in OCD patients, and successful treatment tends to normalise this hyperactivity. This aligns perfectly with a predictive coding account: the OFC/ACC could be considered hubs of computing whether outcomes match predictions (ACC signals "error/conflict," OFC encodes expected value and violation). Their hyperactivity means the brain is constantly flagging potential errors or unexpected outcomes, even when unwarranted.

Computational models suggest that increasing the precision of certain prediction errors can produce OCD-like repetitive checking behaviour as the system tries to resolve an endless stream of perceived mismatches [52, 50]. Furthermore, recent theoretical work proposes OCD as a disorder of hyper-attention to internal signals – effectively, too much attention (precision) is allocated to thoughts of potential disaster, and not enough to external evidence that things are okay.

4.5 Anxiety Disorders: Exaggerated Uncertainty and Threat Predictions

Anxiety disorders, including generalised anxiety and panic disorder, can be viewed through predictive coding as disorders of threat inference under uncertainty. Anxious individuals often overestimate the likelihood of harm and have difficulty tolerating uncertainty. In predictive processing terms, there may be an over-precision [55, 45, 56] of negative predictions about the future, coupled with an inability to down-regulate error signals related to ambiguous stimuli. The result is a chronic state of expecting the worst (a strongly weighted prior of threat) and a hypersensitivity to any signal that might indicate danger (even if it's equivocal).

One formal expression is that the estimated volatility of the environment is high – the brain assumes that unexpected bad events can happen at any time, thus it keeps prediction errors for threat cues high and doesn't easily extinguish fear expectations. A simple model might add a noise term ω to state transitions and treat it as very large in anxiety:

$$x_{t+1} = f(x_t) + \omega_t \tag{8}$$

with the anxious brain assuming large ω (environmental uncertainty). This leads to maintaining a state of vigilance as the optimal strategy. While sharing some features with OCD, anxiety disorders are distinct in their predictive coding profile. Where OCD involves excessive precision on specific error signals (e.g., "the door might not be locked"), anxiety disorders feature a more generalized overestimation of environmental volatility and threat probability. Different anxiety subtypes may represent variations within this framework: generalized anxiety disorder involves broad overestimation of threat across multiple domains, panic disorder features catastrophic misinterpretations of bodily sensations as immediate threats, and specific phobias exhibit localized precision abnormalities for particular stimuli [17].

Neuroimaging findings in anxiety align with this picture. The amygdala, a key region for threat processing, shows heightened activation to uncertain or ambiguous cues in anxious individuals [57]. Essentially, the amygdala responds with alarm even when a situation is only potentially aversive, reflecting a failure to attenuate prediction errors about threat in uncertain contexts. At the same time, top-down regulation from the prefrontal cortex (which in predictive coding would convey reassuring predictions or reappraisals) is often weaker in anxiety. This combination – strong bottom-up error signals for threat, weak top-down calming predictions – yields a dominance of "fear prediction errors" that sustain anxiety [58].

Clinically, this maps to phenomena like hyper-vigilance (constantly scanning for danger) and intolerance of uncertainty (distress when outcomes are unpredictable). In computational terms, one could say the prior for a safe outcome has abnormally low precision in anxious individuals; instead, the prior expectancy might be biased toward danger, and any deviation (even safe signals) fails to fully convince the system that all is well. Treatments like exposure therapy can be seen as attempts to recalibrate these prediction weights – by repeated safe exposures, the patient's brain is encouraged to assign greater weight to the "I am safe" prediction relative to the "something bad will happen" prediction error [59].

4.6 Bipolar Disorder: Instability of Precision and Network Dynamics

Bipolar disorder, marked by oscillations between manic (or hypo-manic) and depressive states, can be conceptualised in predictive coding terms as an instability in how the brain regulates precision across different states [43, 47]. One proposal is that bipolar disorder involves difficulty maintaining a consistent hierarchical inference: the neurotransmitter and modulatory systems that set precision (like NMDA, dopamine and serotonin) fluctuate abnormally, causing the brain to over-fit at times and under-fit at others [60].

$$\frac{d\mu}{dt} = \alpha(t) \cdot \varepsilon \cdot \frac{\partial \hat{y}}{\partial \mu} \tag{9}$$

Here, μ denotes an internal belief or *mood-related latent state*, ε is the prediction error, and $\alpha(t)$ is a time-varying precision or gain parameter (modulated by NMDA, dopamine or serotonin) [43, 47]. In bipolar disorder, $\alpha(t)$ becomes dysregulated - increasing excessively in mania (leading to overconfident updates and reward-seeking behaviour), and dropping in depression (blunting belief updates and reinforcing negative expectations). This instability pushes the system between maladaptive attractor states, resulting in oscillations in mood and behaviour.

In mania, the brain may assign excessive precision to active, exploratory policies and reward-predicting priors ("everything will turn out great"), leading to overconfidence, racing thoughts, and risk-taking (the internal model is too strongly believed). In depression (the other pole), the precision might crash for positive priors, leading to the state described earlier of hypo-learning and negativity [61, 62]. Thus, the homeostatic control of precision weights fails to keep the system in balance, and it instead switches between attractor states of high vs. low confidence. One can imagine a double-well energy landscape for brain states – one basin corresponds to depressive mode, another to manic mode – and the system unpredictably jumps between them due to regulatory noise.

State-space models have been used to simulate such phenomena, where a parameter representing gain or arousal varies over time and pushes the system from one regime to another [1]. For instance, a simplified model might have: when precision parameter α is above a threshold, the network engages a "manic" pattern of activity (high reward-seeking, low error sensitivity), and when α falls below a threshold, a "depressive" pattern emerges. Abrupt neuromodulatory shifts (say, in dopamine tone) could trigger these transitions.

Neuroimaging studies of bipolar disorder have indeed found differences in connectivity between mood states – e.g. mania is associated with heightened connectivity in reward circuits and reduced prefrontal oversight, whereas depression shows the opposite. There is evidence of dysregulated oscillatory activity and signalling between the prefrontal cortex and limbic regions (like the amygdala) in bipolar patients. In predictive coding terms, this might reflect inconsistent application of top-down constraints on emotional inference: sometimes too much (leading to an inflated, unfettered mood in mania) and sometimes too little (leading to depressive pessimism) [62, 63].

While bipolar disorder is complex and not as extensively modelled in predictive coding as other disorders, this perspective provides a coherent narrative: it is a disorder of regulatory oscillation, where the mechanisms that normally stabilise our predictive mind – keeping emotional predictions attuned to reality – themselves become unstable.

5 Discussion: Implications for Psychiatry

The above examples illustrate how predictive coding can unify our understanding of diverse psychiatric symptoms under a common computational framework. This approach may have several broad implications for research and clinical practice, which we explore in the following subsections.

5.1 Mechanistic Biomarkers

Embracing computational models opens the door to identifying biomarkers based on circuit function rather than just phenomenology [64, 65, 66]. Instead of purely descriptive diagnoses, clinicians could measure specific deviations in a patient's predictive coding dynamics (for example, an EEG marker of abnormal error signalling or a connectivity pattern from DCM) as a biomarker of illness. Such biomarkers would directly reflect underlying mechanisms – for instance, reduced top-down connectivity or heightened sensory precision – linking symptoms to neurobiology [27].

This is in line with the goals of precision psychiatry, which seeks objective, quantitative measures to classify and treat mental disorders. Computational psychiatry is seen as an essential tool in this effort, helping to translate between observed behaviour/neural data and the latent neurocomputational parameters that differ across individuals. Over time, a catalogue of predictive coding abnormalities (e.g. "hyper-precision of threat prediction" for certain anxiety disorders, or "NMDA hypofunction leading to reduced error correction" for certain psychoses) could form a basis for a new nosology grounded in mechanism. Importantly, these models can be iteratively refined and validated against longitudinal data, improving their reliability as biomarkers.

5.2 Personalised and Precision Treatment

A computational perspective can inform treatment by tailoring interventions to the patient's specific predictive processing profile. For instance, two patients might both have anxiety, but one might show exaggerated bottom-up error signals (sensory hypervigilance) while another shows mainly a top-down prior bias (catastrophic thinking). These nuances could suggest different treatments [60]: perhaps the former would benefit more from stimulus-driven desensitisation (to recalibrate error responses), whereas the latter might benefit from cognitive restructuring techniques (to adjust overly precise priors).

On the pharmacological side, if a model indicates that a patient's symptoms stem from low precision in a certain circuit (implying maybe inadequate neuromodulatory drive), one might choose a drug that boosts that neuromodulator [1]. In schizophrenia, for example, the predictive coding account implicates dopamine in precision control; treatments that restore dopamine balance (antipsychotics) can be understood as partially re-tuning precision weights to normal levels. More broadly, *in silico* modelling can simulate how a given patient's brain might respond to different interventions. This aligns with the concept of personalised psychiatry, using a patient's data in a model to predict the optimal treatment strategy. While still in early stages, such approaches could improve outcomes by moving beyond one-size-fits-all therapy toward individualised care plans based on computational phenotyping.

Along these lines, predictive coding also provides a framework upon which to integrate psychotherapy and behavioural treatments. With a given disorder or dysfunction described in biologically informed predictive coding terms it becomes possible to map the contribution of a drug (such as ketamine for depression) to alleviating depression (pharmacologically increasing sensitivity to prediction errors as well as imposing a more positive attractor state) and the usefulness of therapy to maximise this opportunity such as by working to imbed the "corrected" thinking to maintain the

healthy state. Ketamine has indeed been shown to increase sensitivity to prediction error in the hours where the antidepressant state emerges [67]. Antidepressant medicines have been shown to be most effective when combined with therapy.

5.3 Novel Therapeutics and Interventions

The predictive coding framework inspires new intervention approaches as well. One exciting area is computationally informed neurostimulation. For example, if OCD is conceptualised as a hyperactive error signal in ACC, treatments like deep brain stimulation (DBS) could be guided to specifically down-regulate that error unit activity. There are efforts to design closed-loop stimulation devices that use real-time recordings to adjust stimulation in response to abnormal neural patterns (like a burst of pathological error signalling). In theory, a closed-loop system could continuously drive the brain toward a lower free-energy state, essentially helping the patient's brain to more effectively minimise prediction errors.

Another avenue is training paradigms or biofeedback: patients could be given tasks that implicitly rebalance their predictive coding [52]. For instance, video games that reward the patient for reinterpreting surprising cues might strengthen certain neural pathways for error processing. Moreover, the emphasis on active inference (the idea that the brain not only passively updates but also takes actions to fulfil predictions) suggests behavioural interventions could aim to break maladaptive active inference loops [68]. For example, encouraging patients with depression to engage in new, surprising activities can provide prediction errors that force an update to their negative priors (essentially "shaking" the system out of the depressive attractor).

Finally, pharmacotherapy development can benefit from these models by targeting the identified circuit parameters: drugs that affect synaptic gain, adaptation, or oscillatory coupling might be tested in computational models of disorders before clinical trials, increasing the rationale for certain targets. In sum, predictive coding offers a principled framework to design interventions that steer the brain's computations, whether through chemicals, devices, or behavioural experience, to restore healthy inference.

5.4 Challenges

Despite its promise, translating computational psychiatry from theory to clinical practice faces several key challenges. Foremost among them is the difficulty of integrating computational insights into clinical taxonomies and diagnostic systems in a way that allows for sensitive and specific identification of pathology.

One major issue is *co-morbidity*. For instance, autism and schizophrenia are often comorbid, yet they are frequently cited as archetypal examples of contrasting predictive coding mechanisms. Any clinically useful tool must be capable of detecting both conditions when they co-occur in the same individual. The fact that they can co-occur implies that predictive coding alterations in the brain may exhibit some degree of domain specificity, and any diagnostic framework must be designed accordingly.

Another challenge is the *marked heterogeneity within diagnostic categories*, particularly mood disorders. Two individuals may receive the same diagnosis-such as major depressive disorder or schizophrenia-yet share no overlapping symptoms. Anhedonia experienced in bipolar disorder has been found to have partially distinct neurophysiology to that in unipolar depression [69]. Furthermore, diagnoses often evolve over time. A person initially diagnosed with major depressive disorder may later be reclassified as bipolar following the emergence of mania, and may subsequently transition from bipolar type II to type I. Similarly, an individual experiencing cyclical mood changes might initially receive a diagnosis of bipolar disorder, only to have it later revised to premenstrual dysphoric disorder when the temporal pattern of symptoms is recognised.

These examples highlight a critical point: any attempt to ground neurocomputational psychiatry in mechanistic frameworks must not inherit the foundational limitations of the The Diagnostic and Statistical Manual of Mental Illnesses (DSM) [70]. Diagnostic instability and symptom heterogeneity challenge the assumption that current categories map onto distinct underlying neurobiology.

Indeed, these limitations may help explain why biomarkers grounded in predictive coding-such as mismatch negativity-have not yet demonstrated sufficient specificity or reliability for clinical translation. This remains true whether such biomarkers are assessed using basic DCM or more conventional event related potential analyses. One potential solution is to initially validate these methods using enriched sampling strategies that focus on "ideal" or exemplar participants. However, for true clinical utility, these tools must ultimately prove robust in the messiness of real-world

clinical populations.

These challenges, however, are not fatal to the feasibility of the approach. Although the DSM classifies mental disorders into discrete categories, treatments are often based on symptom domains, which themselves cut across diagnostic boundaries. For example, antipsychotic medications are used to treat psychosis regardless of whether it arises in the context of schizophrenia or bipolar disorder. Similarly, depressed mood across bipolar disorder, premenstrual dysphoric disorder, and major depressive disorder is commonly treated with selective serotonin reuptake inhibitors (SSRIs). While the choice and dosing of medication may be partially informed by overarching diagnostic labels, this is a far cry from the categorical logic of treating a viral infection with an antibiotic - a treatment that is not only ineffective, but biologically inappropriate and without mechanistic rationale. Optimistically speaking, predictive coding may support the development of precision psychiatric medicine meaning optimal choices may be based upon an individuals neurophysiology as well as symptoms (again rather than DSM classification).

In this light, predictive coding may find its greatest utility not in redefining diagnostic categories, but in providing mechanistic descriptions of symptom classes-dimensions of dysfunction that are either necessary for a diagnosis or that support meaningful sub-classification within a diagnosis. Between sub-classifications or even individuals it may support the selection of mechanistically distinct treatments (such as ketamine over and SSRI for depression). Such models offer hypotheses that are mathematically constrained and amenable to empirical testing via neurophysiology.

Another major challenge is *ensuring that the computational parameters used in these models can be reliably estimated* in individual patients using accessible technologies. Advances in electroencephalography (EEG), including increased portability and affordability, offer a promising path forward. Similarly, developments in magnetoencephalography (MEG), particularly the use of optically pumped magnetometers, are making high-resolution, non-invasive brain measurements more feasible in clinical contexts. Surface encephalography, in particular, provides the most accessible, non-invasive route to probing neural circuit dynamics at the level inferred by neural mass and mean field models.

A final challenge concerns *the need for longitudinal validation.* For computational biomarkers to be clinically useful, their stability and predictive value over time must be established. Few studies have examined the test-retest reliability of DCM-based biomarkers or the consistency of individual parameter "fingerprints" across multiple sessions. 6803

6 Conclusion

Predictive coding has quickly become a leading theoretical framework in cognitive and computational neuroscience, and shows great promise in psychiatry. By viewing mental disorders as disturbances in the brain's predictive machinery, we gain a common language to describe phenomena as varied as hallucinations, anxiety, and cognitive inflexibility. This approach encourages researchers and clinicians to think in terms of circuits and computations - how is the brain weighting its predictions vs. errors? which connections are failing to convey predictions? - rather than solely in terms of symptoms and subjective reports. The advantage of such a framework is not only its explanatory power, but also its generality: it leads to testable hypotheses and models that can be validated with neural data. As we refine these models (through approaches like DCM, neural mass modelling, and machine learning on large-scale data), we move closer to identifying the true underlying dimensions of psychopathology that cut across traditional diagnoses. 6955

7 Funding

This work was supported by Wellcome [226709/Z/22/Z]

References

[1] Read Montague, Raymond J Dolan, Karl J Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, 2012.

[2] Rick A Adams, Klaas Enno Stephan, Harriet R Brown, Chris D Frith, and Karl J Friston. The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4:47, 2013.

[3] Frederike H. Petzschner, Lilian A. E. Weber, Tim Gard, and Klaas E. Stephan. Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis. *Biological Psychiatry*, 82(6):421–430, 2017.

[4] Christophe Gauld, Guillaume Dumas, Éric Fakra, Jérémie Mattout, and Jean-Arthur Micoulaud-Franchi. Les trois cultures de la psychiatrie computationnelle. *Annales Médico-psychologiques, revue psychiatrique*, 179(1):63–71, 2021.

[5] Rick A Adams, Quentin JM Huys, and Jonathan P Roiser. Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery, and Psychiatry*, 87(1):53–63, 2016.

[6] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456):815–836, 2005.

[7] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[8] Karl Friston and Klaas Enno Stephan. Free-energy and the brain. *Synthese*, 159(3):417–458, 2007.

[9] Philipp Sterzer, Rick A Adams, Paul Fletcher, Chris Frith, Stephen M Lawrie, Lars Muckli, Predrag Petrovic, Peter Uhlhaas, Martin Voss, and Philip R Corlett. The predictive coding account of psychosis. *Biological Psychiatry*, 84(9):634–643, 2018.

[10] Karl J Friston, Laurel Harrison, and Will Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

[11] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, 2004.

[12] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[13] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[14] Jakob Hohwy. *The Predictive Mind*. Oxford University Press, 2013.

[15] Gustavo Deco, Viktor Jirsa, and Anthony R McIntosh. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Computational Biology*, 4(8):e1000092, 2008.

[16] Michael Breakspear. Dynamic models of large-scale brain activity. *Nature Neuroscience*, 20(3):340–352, 2017.

[17] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[18] Rosalyn J Moran, Dimitris A Pinotsis, and Karl Friston. Neural masses and fields in dynamic causal modeling. *Frontiers in Computational Neuroscience*, 7:57, 2013.

[19] Ben H Jansen and Vincent G Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics*, 73(4):357–366, 1995.

[20] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.

[21] Rosalyn J Moran, Klaas E Stephan, Thomas Seidenbecher, Hans-Christian Pape, Raymond J Dolan, and Karl J Friston. Consistent spectral predictors for dynamic causal models of steady-state responses. *NeuroImage*, 55(4):1694–1708, 2011.

[22] Klaas Enno Stephan, Karl J Friston, and Christopher D Frith. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin*, 35(3):509–527, 2008.

[23] Lioba C S Berndt, Krish D Singh, and Alexander D Shaw. Restoring synaptic balance in schizophrenia: Insights from a thalamo-cortical conductance-based model. *bioRxiv*, 2024. Preprint.

[24] Nikola T Markov, Julien Vezoli, Pascal Chameau, et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259, 2014.

[25] S. Murray Sherman and R. W. Guillery. *Functional connections of cortical areas: A new view from the thalamus*. MIT Press, 2013.

[26] Karl J. Friston, Katrin H. Preller, Christoph D. Mathys, Hayriye Cagnan, Joachim Heinzle, Adeel Razi, and Peter Zeidman. Dynamic causal modelling revisited. *NeuroImage*, 199:730–744, 2019.

[27] Karl Friston, A David Redish, and Jeremy A Gordon. Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2):148–158, 2014.

[28] Philip R Corlett, Jennifer A Mollick, Hedy Kober, et al. Hallucinations and strong priors. *Trends in Cognitive Sciences*, 23(2):114–127, 2019.

[29] Klaas Enno Stephan and Christoph Mathys. Computational approaches to psychiatry. *Current Opinion in Neurobiology*, 25:85–92, 2014.

[30] Philip R. Corlett, Garry D. Honey, and Paul C. Fletcher. Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology*, 30(11):1145–1155, Nov 2016.

[31] Alexander D Shaw, Laura Knight, Tom CA Freeman, Gemma M Williams, Rosalyn J Moran, Karl J Friston, James TR Walters, and Krish D Singh. Oscillatory, computational, and behavioral evidence for impaired gabaergic inhibition in schizophrenia. *Schizophrenia Bulletin*, 46(2):345–353, 2020.

[32] Daniel Umbricht and Sasa Krljes. Mismatch negativity in schizophrenia: a meta-analysis. *Schizophrenia Research*, 76(1):1–23, 2005.

[33] Albert R Powers, Christoph Mathys, and Philip R Corlett. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351):596–600, 2017.

[34] Anil K Seth and Karl J Friston. Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B*, 371(1708):20160007, 2016.

[35] Elizabeth Pellicano and David Burr. When the world becomes 'too real': a bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10):504–510, 2012.

[36] Sander Van de Cruys, Kasper Evers, Ruth Van der Hallen, Lise Van Eylen, Bart Boets, Lisa de Wit, and Johan Wagemans. Precise minds in uncertain worlds: predictive coding in autism. *Psychological Review*, 121(4):649–675, 2014.

[37] Rebecca P. Lawson, Geraint Rees, and Karl J. Friston. An aberrant precision account of autism. *Frontiers in Human Neuroscience*, 8:302, 2014.

[38] Francesca Happé and Uta Frith. The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of autism and developmental disorders*, 36:5–25, 2006.

[39] Elizabeth Pellicano and David Burr. When the world becomes 'too real': a bayesian explanation of autistic perception. *Trends in cognitive sciences*, 16(10):504–510, 2012.

[40] John L R Rubenstein and Michael M Merzenich. Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes, Brain and Behavior*, 2(5):255–267, 2003.

[41] Vikaas S Sohal and John L R Rubenstein. Excitatory–inhibitory balance as a framework for investigating mechanisms in neuropsychiatric disorders. *Molecular Psychiatry*, 24(9):1248–1257, 2019.

[42] Ryota Kanai, Yuka Komura, Stewart Shipp, and Karl Friston. Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B*, 370(1668):20140169, 2015.

[43] Quentin JM Huys, Nathaniel D Daw, and Peter Dayan. Depression: a decision-theoretic analysis. *Annual Review of Neuroscience*, 38:1–23, 2015.

[44] Tobias Kube. Biased belief updating in depression. *Clinical Psychology Review*, 103:102298, 2023.

[45] Erdem Pulcu and Michael Browning. A misestimation of uncertainty in affective disorders. *Psychological Medicine*, 49(3):403–411, 2019.

[46] Robb B. Rutledge, Norah Skandali, Peter Dayan, and Raymond J. Dolan. Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry*, 74(7):790–797, 2017.

[47] Peter Dayan and Quentin JM Huys. Serotonin, inhibition, and negative mood. *PLoS Computational Biology*, 4(2):e4, 2008.

[48] Neil S Jacobson, Christopher R Martell, and Sona Dimidjian. Behavioral activation treatment for depression: returning to contextual roots. *Clinical Psychology: science and practice*, 8(3):255, 2001.

[49] Aaron T Beck. *Cognitive therapy and the emotional disorders*. Penguin, 1979.

[50] Isaac Fradkin, Rick A. Adams, Thomas Parr, Jonathan P. Roiser, and Jonathan D. Huppert. Searching for an anchor in an unpredictable world: A computational model of obsessive-compulsive disorder. *Psychological Review*, 127(5):672–699, 2020.

[51] Yi-Jie Zhao, Yingying Zhang, Qianfeng Wang, Luis Manssuer, Hailun Cui, Qiong Ding, Bomin Sun, Wenjuan Liu, and Valerie Voon. Evidence accumulation and neural correlates of uncertainty in obsessive-compulsive disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(10):1058–1065, Oct 2023.

[52] Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: optimal control without cost functions. *Biological Cybernetics*, 106(8):523–541, 2012.

[53] Wenxin Tang, Qifeng Zhu, Xiangyang Gong, Cheng Zhu, Yiquan Wang, and Shulin Chen. Cortico-striato-thalamo-cortical circuit abnormalities in obsessive-compulsive disorder: A voxel-based morphometric and fmri study of the whole brain. *Behavioural Brain Research*, 313:17–22, 2016.

[54] Claire M Gillan and Trevor W Robbins. Goal-directed learning and obsessive–compulsive disorder. *Philosological Transactions of the Royal Society B*, 369(1655):20130475, 2014.

[55] Michael Browning, Timothy EJ Behrens, Gerhard Jocham, Jane X O'Reilly, and Sonia J Bishop. Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, 18(4):590–596, 2015.

[56] Martin P Paulus and Angela J Yu. Emotion and decision-making: affect-driven belief systems in anxiety and depression. *Trends in Cognitive Sciences*, 16(9):476–483, 2012.

[57] Dan W Grupe and Jack B Nitschke. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nature Reviews Neuroscience*, 14(7):488–501, 2013.

[58] Amit Etkin and Tor D Wager. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in ptsd, social anxiety disorder, and specific phobia. *American Journal of Psychiatry*, 164(10):1476–1488, 2007.

[59] Akshay Nair, Robb B. Rutledge, and Liam Mason. Under the hood: Using computational psychiatry to make psychological therapies more mechanism-focused. *Frontiers in Psychiatry*, 11:140, 2020.

[60] Leanne M Williams. Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *The Lancet Psychiatry*, 3(5):472–480, 2016.

[61] Thien-Thao T. Nguyen, Sanja Kovacevic, S. I. Dev, Kun-Chia Lu, T. T. Liu, and Lisa T. Eyler. Dynamic functional connectivity in bipolar disorder is associated with executive function and processing speed: A preliminary study. *Neuropsychology*, 31(1):73–83, 2017.

[62] Mary L Phillips and Holly A Swartz. A critical appraisal of neuroimaging studies of bipolar disorder: toward a new conceptualization of mood disorders. *Biological Psychiatry*, 75(6):434–440, 2014.

[63] Duk-Ju Kim, Ashley R. Bolbecker, Jessica Howell, Ofer Rass, Olaf Sporns, William P. Hetrick, and Dost Öngür. Disturbed resting state eeg synchronization in bipolar disorder: A graph-theoretic analysis. *NeuroImage: Clinical*, 2:414–423, 2013.

[64] Klaas E Stephan, Florian Schlagenhauf, Quentin JM Huys, Suyog Raman, Lorenz Deserno, and Andreas Heinz. Translational perspectives for computational neuroimaging. *Neuron*, 87(4):716–732, 2015.

[65] Quentin JM Huys, Tiago V Maia, and Michael J Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, 2016.

[66] Stefan Frässle, Yunzhe Yao, Daniel Schöbi, Eduardo Aponte, Jakob Heinzle, and Klaas E Stephan. Generative models for clinical applications in computational psychiatry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(3):e1560, 2022.

[67] Rachael L. Sumner, Rebecca McMillan, Meg J. Spriggs, Doug Campbell, Gemma Malpas, Elizabeth Maxwell, Carolyn Deng, John Hay, Rhys Ponton, Frederick Sundram, and Suresh D. Muthukumaraswamy. Ketamine improves short-term plasticity in depression by enhancing sensitivity to prediction errors. *European Neuropsychopharmacology*, 38:73–85, 2020.

[68] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. Active inference: the free energy principle in mind, brain, and behavior. *MIT Press*, 2022.

[69] Alexis E. Whitton and Diego A. Pizzagalli. *Anhedonia in Depression and Bipolar Disorder*, pages 111–127. Springer International Publishing, Cham, 2022.

[70] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed., text rev.)*. American Psychiatric Publishing, 2022.