JNphi

Predictive Processing and Active Inference: A Comprehensive Review of Theoretical Foundations, Neural Mechanisms, and Clinical Implications in Cognitive Science

Taruna Ikrar, Wachyudi Muchsin, Alfi Sophian*

Abstract

Predictive processing (PP) and its active counterpart, active inference (AI), have emerged as among the most influential frameworks in contemporary cognitive science, offering a unified account of perception, cognition, action, and learning under a single computational principle: the minimization of prediction error or, more formally, the minimization of variational free energy. Originally grounded in Helmholtzian notions of perception as unconscious inference, the framework has been substantially formalized through Karl Friston's free energy principle (FEP), which proposes that all self-organizing biological systems resist entropy by implicitly minimizing surprise. This review critically examines the theoretical underpinnings of the PP/AI framework, its neurobiological implementation via hierarchical predictive coding in cortical circuits, and its behavioral and clinical manifestations across a spectrum of psychological conditions. We synthesize evidence from computational modeling, electrophysiology, neuroimaging, and psychophysics to evaluate the empirical status of the framework. We further discuss unresolved controversies—including the explanatory scope of the FEP, the interpretation of precision-weighting, and the relationship between PP and alternative computational theories of mind—and identify promising directions for future research. We conclude that while PP/AI represents a transformative theoretical advance, its full explanatory power remains contingent on tighter integration with mechanistic neuroscience and rigorous empirical testing.

Key Words: predictive processing; active inference; free energy principle; hierarchical predictive coding; Bayesian brain

DOI: 10.5281/zenodo.20049615

Corresponding author: Alfi Sophian
Address: Indonesia FDA, Jl. Percetakan Negara, No.23, Jakarta Pusat, 10560, Indonesia
e-mail ✉ alfi.sophian@pom.go.id

1. INTRODUCTION

The question of how the brain constructs coherent representations of the world from inherently ambiguous and noisy sensory data has occupied philosophers and scientists for centuries. From Kant's (1781/1998) transcendental aesthetics to Helmholtz's (1867) notion of perception as 'unconscious inference,' the idea that the mind actively constructs rather than passively receives perceptual experience has a distinguished intellectual lineage. In contemporary cognitive neuroscience, this constructivist tradition finds its most formalized expression in the predictive processing (PP) framework, and its generalization, active inference (AI).

The PP/AI framework proposes that the brain is fundamentally a prediction machine—a hierarchically organized inference engine that continuously generates top-down probabilistic predictions about the causes of sensory signals and updates these predictions based on bottom-up prediction errors (Clark, 2016; Friston, 2010; Rao & Ballard, 1999). Rather than passively relaying sensory data upward to higher cortical areas, the brain is conceived as issuing hypotheses about the world and using incoming signals primarily as a vehicle for updating and refining these hypotheses. This Bayesian perspective on brain function has profound implications not only for understanding normal perception and cognition, but also for explaining psychopathology, consciousness, and the nature of agency.

The formal backbone of the framework—Friston's free energy principle (FEP)—offers a mathematically precise account of how biological systems maintain their organizational integrity in the face of a constantly changing environment (Friston, 2010; Friston et al., 2006). By minimizing variational free energy—a bound on the log-evidence for a generative model of the world—organisms can be understood as implicitly performing approximate Bayesian inference on the hidden causes of their sensory inputs. Crucially, the FEP subsumes action under the same inferential scheme as perception: rather than moving to execute motor commands per se, organisms act to bring sensory data into conformity with their generative models, a process termed active inference (Friston et al., 2010).

The scope of PP/AI has expanded rapidly since Rao and Ballard's (1999) seminal computational model, encompassing accounts of attention (Feldman & Friston, 2010), emotion and interoception (Seth & Friston, 2016), social cognition (Frith & Frith, 2012), language (Pickering & Garrod, 2013), and psychopathology (Adams et al., 2013; Corlett et al., 2019). This breadth has made PP/AI both influential and controversial: critics question whether the framework's generality renders it unfalsifiable, while proponents argue that its unifying potential is precisely what distinguishes it from narrower computational accounts (see Clark, 2019; Colombo & Seriès, 2012).

The present review aims to provide a critical narrative evaluation of the PP/AI framework. We proceed as follows: Section 2 outlines the theoretical foundations; Section 3 reviews the neurobiological evidence; Section 4 examines behavioral and computational evidence; Section 5 surveys clinical applications; Section 6 addresses critiques and open questions; and Section 7 charts directions for future research.

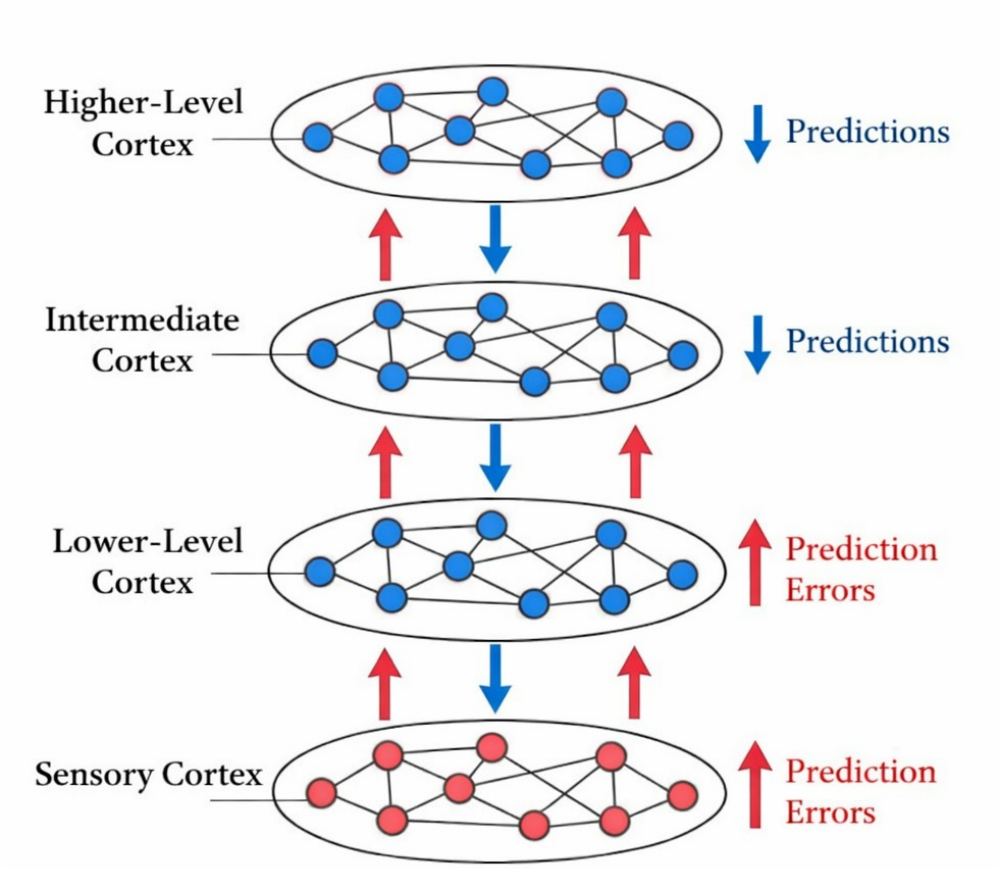

Figure 1. Schematic overview of the hierarchical predictive processing architecture. Top-down predictions (blue arrows) are compared with bottom-up sensory signals (red arrows) at each cortical level, generating prediction errors (PE) that propagate upward. Source: Rao & Ballard (1999). Predictive coding in the visual cortex. Nature Neuroscience, 2(1), 79–87.; Friston (2008). Hierarchical models in the brain. PLoS Computational Biology, 4(11), e1000211.; Bastos et al. (2012). Canonical microcircuits for predictive coding. Neuron, 76(4), 695–711.

2. THEORETICAL FOUNDATIONS

2.1 Helmholtz, Bayes, and the Bayesian Brain

The intellectual origins of predictive processing can be traced to Hermann von Helmholtz's (1867) concept of perception as 'unconscious inference'—the idea that perceptual experience is the product of an implicit inferential process that combines sensory evidence with prior knowledge about the world. This idea lay largely dormant until it was revived and formalized in computational terms by theorists working in the Bayesian tradition (Knill & Pouget, 2004; Weiss et al., 2002).

The Bayesian brain hypothesis proposes that the brain represents and manipulates probability distributions over possible states of the world, combining prior beliefs with likelihood functions derived from sensory evidence to compute posterior beliefs (Körding & Wolpert, 2004). This hypothesis has received considerable empirical support from behavioral studies demonstrating that humans perform near-optimal probabilistic inference across a wide range of perceptual and motor tasks (Ernst & Banks, 2002; Körding & Wolpert, 2004).

Rao and Ballard's (1999) hierarchical predictive coding model provided an influential computational implementation of the Bayesian brain hypothesis for the visual system. In their model, higher cortical areas send predictions about the activity of lower areas via top-down connections, and lower areas return only the residual prediction error—the discrepancy between the prediction and the actual sensory signal—via bottom-up connections. This architecture achieves efficient coding by transmitting only the unexpected components of sensory signals, a principle consistent with the principle of efficient coding proposed by Barlow (1961).

2.2 The Free Energy Principle

Karl Friston's free energy principle (FEP) provides the most mathematically comprehensive formulation of the predictive processing framework (Friston, 2010; Friston et al., 2006). Grounded in variational Bayesian inference and information theory, the FEP proposes that all self-organizing biological systems—from single cells to entire organisms—resist the tendency toward disorder (entropy) by minimizing a quantity called variational free energy.

Formally, variational free energy (F) is defined as:

$$F = E_q[\ln q(x) - \ln p(o, x)]$$

where $q(x)$ is the variational density (the organism's beliefs about hidden states x), $p(o, x)$ is the generative model (encoding the organism's beliefs about how hidden states give rise to observations o), and E_q denotes expectation under q. Crucially, free energy is an upper bound on surprise (or negative log-evidence): $F \geq -\ln p(o)$. This means that by minimizing free energy, organisms implicitly minimize

surprise—they act and perceive so as to inhabit expected, self-consistent sensory states (Friston, 2010).

The FEP subsumes prediction error minimization as a special case: under Gaussian assumptions, minimizing free energy is equivalent to minimizing precision-weighted prediction errors across all levels of a cortical hierarchy (Friston, 2008). The precision-weighting mechanism is of particular importance, as it allows the framework to account for attentional phenomena: attention is understood as the optimization of precision—the inverse variance of prediction errors—such that more precise (reliable) prediction errors exert greater influence on belief updating (Feldman & Friston, 2010).

2.3 Active Inference and the Control of Action

A distinctive feature of the PP/AI framework is its treatment of action as a form of inference (Friston et al., 2010). In classical predictive processing accounts, prediction errors arise when the world fails to conform to the brain's predictions. The brain can reduce these errors in two ways: (1) by updating its internal model (perception/learning), or (2) by acting on the world to bring it into conformity with its predictions (action). This latter strategy—acting to fulfill predictions—is the core of active inference.

Active inference reconceptualizes action as the fulfillment of proprioceptive and interoceptive predictions generated by high-level goals or preferred states (Friston et al., 2010; Friston et al., 2017). Rather than computing an optimal motor command, the motor system acts to minimize the discrepancy between predicted and actual bodily states. This account has several appealing features: it dissolves the classical distinction between perception and action, explains why organisms with detailed internal models of their own bodies can perform flexible, skilled behaviors, and provides a natural account of motor learning and adaptation.

The active inference framework has been further extended to include planning as inference (Friston et al., 2016; Parr & Friston, 2019). In this account, organisms plan by mentally simulating possible future action sequences and selecting those that minimize expected free energy—a quantity that trades off expected surprise (epistemic value) against divergence from preferred outcomes (pragmatic value). This formulation yields a rich account of goal-directed behavior, exploration-exploitation trade-offs, and curiosity-driven learning.

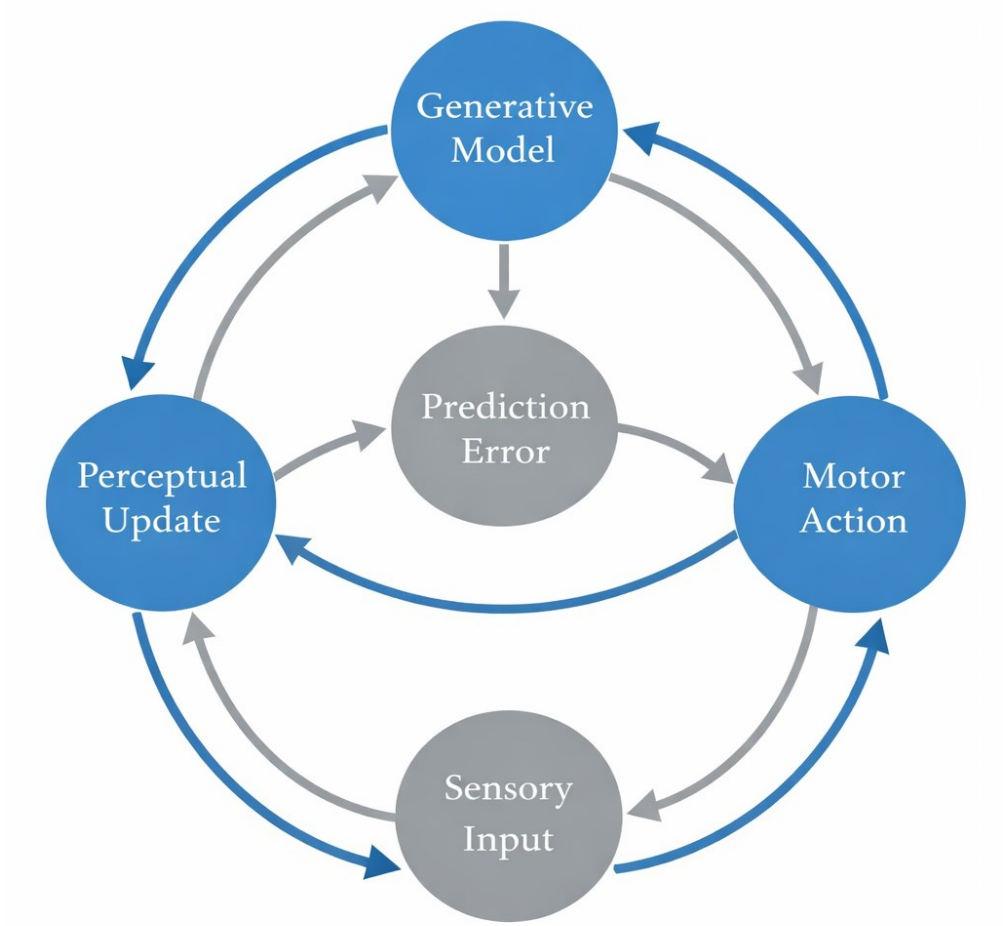

Figure 2. The active inference cycle. Organisms minimize variational free energy through two complementary routes: updating internal beliefs (perceptual inference) or acting to change sensory inputs (active inference). Arrows depict the bidirectional flow between generative model, sensory data, and action. Source: Friston et al. (2010). Action and behavior: A free-energy formulation. Biological Cybernetics, 102(3), 227–260.; Friston et al. (2017). Active inference: A process theory. Neural Computation, 29(1), 1–49.; Parr & Friston (2019). Generalised free energy and active inference. Biological Cybernetics, 113(5–6), 495–513.

3. NEUROBIOLOGICAL EVIDENCE

3.1 Hierarchical Predictive Coding in Cortical Circuits

The predictive coding architecture makes specific claims about the functional organization of cortical circuits that can be tested with neurophysiological methods. According to the framework, top-down connections from higher to lower cortical areas carry predictions, while bottom-up connections carry prediction errors. Furthermore, prediction errors should be encoded by superficial pyramidal neurons,

while predictions are encoded by deep pyramidal neurons, reflecting the known laminar structure of cortical feedback and feedforward projections (Bastos et al., 2012; Friston, 2008).

Bastos et al. (2012) provided an influential synthesis of these predictions, arguing that the laminar specificity of cortical projections—with feedforward projections originating primarily in supragranular layers and feedback projections targeting infragranular and supragranular layers—maps naturally onto the predictive coding scheme. Consistent with this account, gamma-band oscillations (30–100 Hz) have been associated with feedforward (prediction error) signals, while alpha/beta-band oscillations (8–30 Hz) have been associated with feedback (prediction) signals (Bastos et al., 2015; van Kerkoerle et al., 2014).

Electrophysiological studies in humans and non-human primates have provided direct evidence for prediction error signals in cortical areas. The mismatch negativity (MMN)—an event-related potential elicited by unexpected deviations in a regular auditory sequence—is widely interpreted as a neural correlate of auditory prediction error (Näätänen et al., 2001; Garrido et al., 2009). Similar prediction error signals have been identified in the visual, somatosensory, and olfactory systems, suggesting that predictive coding is a general organizing principle of cortical computation (Egner et al., 2010; Kok et al., 2012).

3.2 Precision-Weighting and Neuromodulation

The concept of precision-weighting—the modulation of prediction error gain by estimated reliability—is central to the PP/AI framework's account of attention and learning. Neurobiologically, precision is hypothesized to be encoded by the postsynaptic gain of pyramidal neurons, which in turn is regulated by neuromodulatory systems, particularly acetylcholine and dopamine (Friston et al., 2012; Yu & Dayan, 2005).

Acetylcholine has been proposed to signal expected uncertainty—increasing the precision of likelihood signals relative to priors when sensory environments are volatile (Yu & Dayan, 2005). Dopamine, by contrast, has been associated with signaling precision-weighted prediction errors in the context of reward learning, providing a mechanistic link between the FEP and influential accounts of dopaminergic function in reinforcement learning (Friston et al., 2012; Schultz et al., 1997). Serotonin has been further implicated in temporal discounting and the precision of beliefs about future states (Dayan & Huys, 2009).

Pharmacological and lesion studies provide partial support for these hypotheses. Cholinergic enhancement via acetylcholinesterase

inhibitors has been shown to improve performance on tasks requiring the integration of new sensory evidence over established priors (Vossel et al., 2014), while dopamine receptor blockade disrupts reward prediction error signaling as indexed by both behavior and neural activity (Pessiglione et al., 2006).

3.3 Neuroimaging Evidence

Functional neuroimaging studies have provided converging evidence for the PP/AI framework. A consistent finding is that neural responses to expected stimuli are attenuated relative to unexpected stimuli—a phenomenon known as 'repetition suppression' or 'expectation suppression'—consistent with the framework's prediction that prediction errors, not predictions themselves, drive upward neural activity (Egner et al., 2010; Kok et al., 2012; Summerfield & de Lange, 2014).

Structural equation modeling and dynamic causal modeling (DCM) studies have attempted to characterize the effective connectivity patterns underlying predictive coding, with results broadly consistent with the theoretical framework. For instance, Garrido et al. (2008) used DCM to show that the MMN is better explained by a hierarchical model with both bottom-up and top-down connections than by a purely bottom-up model, supporting the notion of bidirectional prediction and error signals in auditory cortex.

Interoceptive predictive coding—the application of PP principles to bodily signals arising from internal organs—has received increasing attention as a framework for understanding emotion and self-perception. Neuroimaging studies have implicated the anterior insular cortex (AIC) as a key node for integrating interoceptive predictions with visceral signals, with AIC responses tracking the mismatch between expected and actual heartbeat timing (Seth & Friston, 2016; Paulus & Stein, 2006).

4. BEHAVIORAL AND COMPUTATIONAL EVIDENCE

4.1 Perceptual Inference and Illusions

One of the most compelling sources of evidence for the PP/AI framework comes from the study of perceptual illusions and bistable percepts. The framework predicts that perceptual experience reflects the brain's best inference—its maximum a posteriori estimate of the causes of sensory signals—rather than a veridical representation of the external world. Accordingly, illusions arise when prior beliefs or higher-level predictions override or bias the interpretation of sensory data.

The rubber hand illusion (Botvinick & Cohen, 1998) offers a striking example. When participants see a rubber hand stroked in synchrony with their own hidden hand, they report feeling touch on the rubber hand—a manifestation of predictive integration in which visual and tactile prediction errors are jointly minimized by attributing ownership to the visible, rubber hand. Active inference models of this phenomenon capture both the phenomenological and behavioral aspects of the illusion, including the characteristic proprioceptive drift toward the rubber hand (Apps & Tsakiris, 2014).

Bistable perception—phenomena such as the Necker cube or binocular rivalry—are elegantly accounted for within the PP framework as a consequence of ongoing model updating in the face of perceptual ambiguity. When a single sensory signal is equally consistent with two hypotheses, the brain alternates between them, with each alternation corresponding to a switch in the hypothesis that currently best explains the data (Hohwy et al., 2008). Computational models implementing this dynamics have successfully reproduced the statistical properties of perceptual alternation, including the characteristic gamma distribution of percept durations.

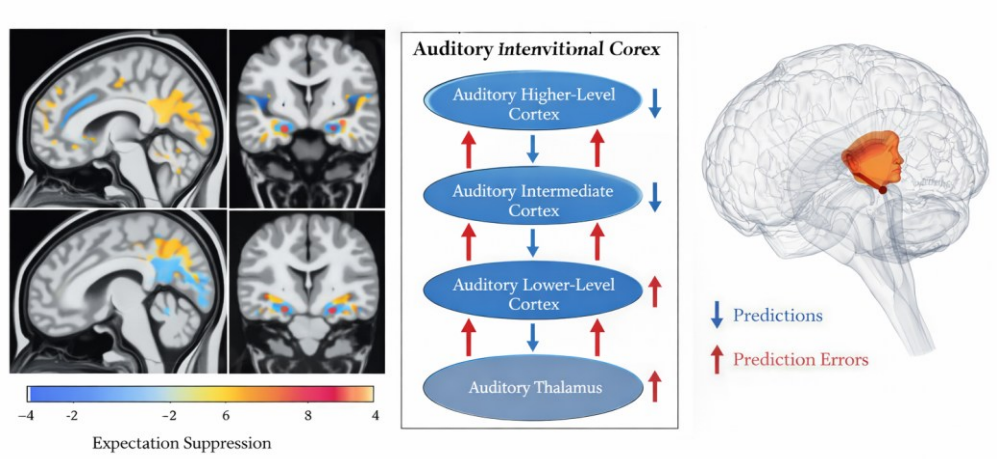

Figure 3. Neuroimaging correlates of predictive processing. (A) Regions showing expectation suppression across sensory modalities. (B) Dynamic causal modeling architecture supporting hierarchical predictive coding in auditory cortex. (C) Anterior insular cortex activation during interoceptive prediction error. Source: Kok et al. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. Neuron, 75(2), 265–270.; Garrido et al. (2009). The mismatch negativity: A review of underlying mechanisms. Clinical Neurophysiology, 120(3), 453–463.; Seth & Friston (2016). Active interoceptive inference and the emotional brain. Philosophical Transactions of the Royal Society B, 371(1708), 20160007.

4.2 Learning and Adaptation

The PP/AI framework offers a unified account of perceptual learning and adaptation: both are understood as the modification of the generative model to reduce long-term prediction error. This

perspective aligns naturally with Hebbian and predictive Hebbian learning rules, and has been used to derive formal models of classical conditioning, sensory adaptation, and skill acquisition (Friston, 2008).

Computational modeling studies have demonstrated that active inference agents can acquire complex, context-appropriate behaviors from minimal experience by combining model-based (epistemic) and model-free (pragmatic) action selection strategies (Friston et al., 2017). This approach offers a principled account of the exploration-exploitation dilemma that is grounded in the same inferential framework as perception and action, rather than requiring separate algorithmic components.

4.3 Social Cognition and Language

The PP/AI framework has been productively extended to the domain of social cognition, where it provides an account of mentalizing (theory of mind) as inference over the hidden mental states of other agents (Frith & Frith, 2012; Kilner et al., 2007). In this view, social perception involves the same hierarchical inferential machinery as object perception, with the difference that the generative model now includes representations of other agents' beliefs, desires, and intentions.

Pickering and Garrod (2013) have proposed a predictive processing account of language comprehension and production, arguing that linguistic communication is fundamentally a process of mutual prediction error minimization between interlocutors. According to this account, language comprehension involves predicting upcoming words and structures, with comprehension difficulty proportional to the prediction error generated by unexpected linguistic input—a prediction that aligns with a large body of psycholinguistic evidence from reading time and neuroimaging studies (Kutas & Federmeier, 2011).

5. CLINICAL IMPLICATIONS

5.1 Computational Psychiatry and the PP Framework

Perhaps the most practically significant extension of the PP/AI framework is its application to understanding and treating mental health conditions—a research program that has come to be known as computational psychiatry (Montague et al., 2012; Stephan et al., 2016). The fundamental insight is that many psychiatric symptoms can be reconceptualized as consequences of aberrant predictive coding: abnormal prior beliefs, miscalibrated precision-weighting, or dysfunctional model updating.

5.2 Schizophrenia Spectrum Disorders

Schizophrenia has been a particularly fruitful domain of application for the PP/AI framework. Adams et al. (2013) and Corlett et al. (2019) have argued that positive symptoms of schizophrenia—hallucinations and delusions—can be understood as consequences of aberrant precision-weighting, specifically an overestimation of sensory prediction error precision relative to prior beliefs. This aberrant precision weighting leads the brain to 'over-explain' prediction errors, generating false inferences about hidden causes that manifest as hallucinations and delusions.

This account aligns with the observation that dopamine dysregulation—a well-established feature of schizophrenia—may disrupt the precision-weighting of prediction errors, since dopamine is hypothesized to modulate the precision of prediction errors in the striatum and prefrontal cortex (Friston et al., 2012; Kapur, 2003). Computational modeling studies have shown that simulated aberrations in precision-weighting reproduce the phenomenological features of psychosis, including the characteristic pattern of 'jumping to conclusions' observed in delusional patients (Moutoussis et al., 2011).

5.3 Anxiety, Depression, and Interoceptive Dysregulation

Anxiety disorders have been conceptualized within the PP framework as a consequence of overly precise prior beliefs about threat, leading to excessive prediction error signaling in response to ambiguous or benign stimuli (Paulus & Stein, 2006; Stein et al., 2011). Active inference accounts of anxiety propose that anxious individuals preferentially select actions that reduce epistemic uncertainty—a computational strategy that manifests as avoidance behavior and intolerance of uncertainty.

Depression, in contrast, has been associated with aberrant precision-weighting that causes excessive reliance on prior beliefs over sensory evidence—a form of 'sticky' prior that leads to pessimistic perceptual bias, diminished reward prediction errors, and reduced plasticity (Stephan et al., 2016; Huys et al., 2015). Interoceptive dysregulation—disrupted precision-weighting of visceral prediction errors—has been proposed as a transdiagnostic mechanism linking anxiety, depression, and somatic symptom disorders (Seth & Friston, 2016; Garfinkel et al., 2015).

5.4 Autism Spectrum Conditions

The PP/AI framework has generated influential accounts of autism spectrum conditions (ASC). Pellicano and Burr (2012) proposed that ASC is characterized by reduced prior beliefs (attenuated priors)

combined with heightened precision of sensory prediction errors—a configuration that would produce unusually detailed sensory experience at the expense of predictive contextual integration. This account aligns with the reported sensory hypersensitivity and reduced susceptibility to perceptual illusions in ASC populations.

Van de Cruys et al. (2014) proposed a complementary account emphasizing the role of aberrant precision in social and sensory contexts, arguing that individuals with ASC have difficulty appropriately scaling the precision of different types of prediction errors—a deficit that manifests as rigidity, restricted interests, and difficulty with social prediction. Computational models implementing these accounts have successfully reproduced patterns of perceptual performance observed in ASC, including reduced susceptibility to the hollow face illusion and altered multisensory integration.

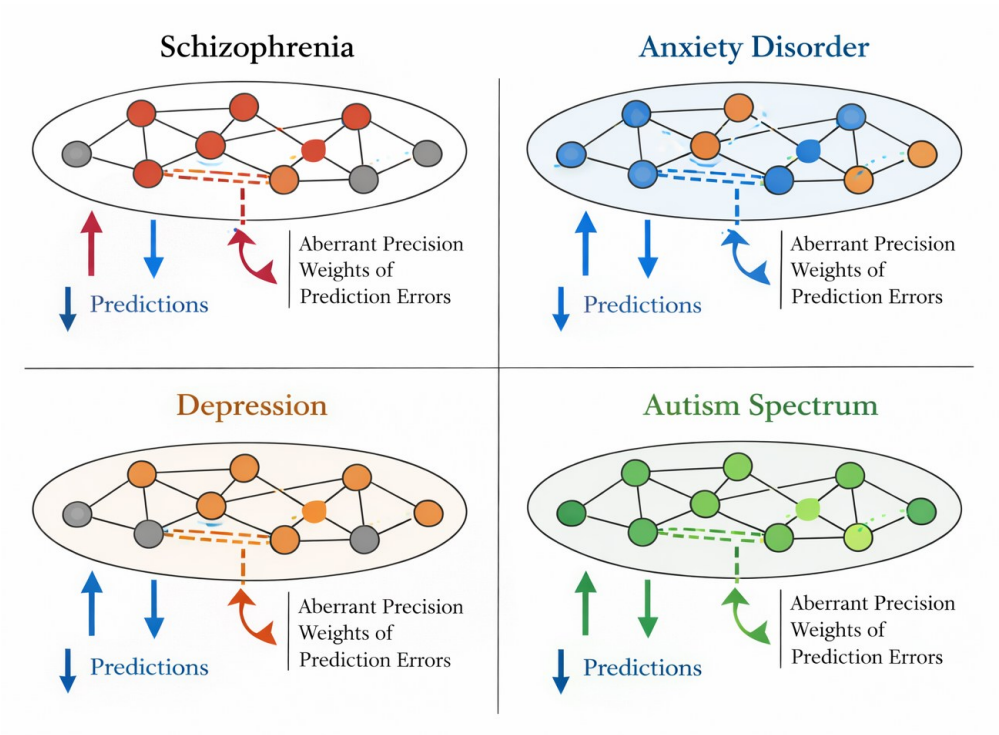

Figure 4. Computational psychiatry applications of the predictive processing framework. Schematic showing how aberrant precision-weighting parameters produce hallmark symptoms across schizophrenia, anxiety, depression, and autism spectrum conditions. Source: Adams et al. (2013). The computational anatomy of psychosis. Frontiers in Psychiatry, 4, 47.; Paulus & Stein (2006). An insular view of anxiety. Biological Psychiatry, 60(4), 383–387.; Pellicano & Burr (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. Trends in Cognitive Sciences, 16(10), 504–510.; Stephan et al. (2016). Charting the landscape of priority problems in psychiatry. The Lancet Psychiatry, 3(1), 77–83.

6. CRITIQUES, CONTROVERSIES, AND OPEN QUESTIONS

6.1 Explanatory Scope and Falsifiability

The PP/AI framework's most prominent critics have raised concerns about its explanatory scope, arguing that the framework's generality risks making it unfalsifiable (Colombo & Seriès, 2012; Colombo, 2017). The FEP, in particular, has been criticized for being a mathematical truism—true by definition for any self-organizing system—rather than a substantive empirical hypothesis (Colombo & Seriès, 2012). From this perspective, the framework cannot be falsified because any behavior can be redescribed as free energy minimization.

Proponents of the framework have responded by distinguishing the mathematical framework itself (which is indeed very general) from the specific process theories and mechanistic models that instantiate it in particular biological systems (Friston et al., 2020). On this view, it is the specific models derived from the FEP—with their concrete computational and neurobiological commitments—that are subject to empirical refutation, even if the overarching framework is not (Hohwy, 2020). This debate reflects a deeper tension in theoretical neuroscience between the pursuit of unifying principles and the requirement for mechanistic specificity.

6.2 The Role of Representation

The PP/AI framework's heavy reliance on internal generative models and probabilistic representations has been challenged by enactivist and ecological accounts of cognition, which emphasize the role of agent-environment dynamics over internal representations (Clark, 2016; Colombo, 2017). Radical enactivists (e.g., Hutto & Myin, 2013) argue that basic cognition does not involve representations at all, and that the PP/AI framework's commitment to generative models re-introduces a problematic Cartesian inner theater.

This debate connects to broader questions about the explanatory role of representation in cognitive science. While some proponents of PP/AI argue that generative models need not be construed as explicit propositional representations (Clark, 2016), others maintain that the computational commitments of the framework are substantively representationalist and that this is a feature, not a bug, of the theory (Hohwy, 2013).

6.3 Integration with Alternative Frameworks

A significant open question concerns the relationship between PP/AI and other influential computational frameworks in cognitive science, including reinforcement learning (RL), predictive coding in its more classical forms, and global workspace theory. While Friston and

colleagues have argued that RL can be derived from the FEP as a special case (Friston et al., 2009), critics note that the relationship is more complex and that the FEP-derived account of reward and goal-directedness may not straightforwardly reduce to standard RL frameworks (Colombo, 2017).

Global workspace theory (Baars, 1988; Dehaene et al., 2003) offers an alternative account of conscious access and the brain-wide integration of information that complements, but does not straightforwardly map onto, the PP/AI framework. Friston et al. (2021) have recently argued that active inference and global workspace theory can be integrated within a common formal framework, but this synthesis remains to be fully elaborated and empirically evaluated.

7. FUTURE DIRECTIONS

Several priority areas emerge from this review. First, tighter integration between the PP/AI framework and mechanistic neuroscience is urgently needed. While theoretical accounts of predictive coding in cortical circuits are increasingly well-developed (Bastos et al., 2012), the mapping from computational-level descriptions to concrete synaptic and circuit-level mechanisms remains incomplete. Multielectrode recording studies combined with optogenetic perturbations offer a promising path toward identifying the neural substrates of prediction, prediction error, and precision-weighting at the cellular level (Keller & Mrsic-Flogel, 2018).

Second, the development of computational phenotyping approaches for clinical populations holds considerable promise. If individual differences in generative model parameters can be reliably estimated from behavioral and neuroimaging data, these parameters may serve as mechanistically interpretable biomarkers for psychiatric diagnosis and treatment response (Montague et al., 2012; Stephan et al., 2016). Longitudinal studies tracking computational phenotypes over the course of illness and treatment are particularly needed.

Third, the application of PP/AI to complex social and cultural phenomena represents a frontier that has thus far received relatively little systematic attention. The extension of active inference to multi-agent settings (Friston et al., 2023) offers formal tools for modeling the emergence of shared norms, communication, and cultural transmission, with potential implications for understanding collective behavior, public health, and institutional design.

Fourth, the integration of the PP/AI framework with advances in machine learning and artificial intelligence—particularly deep generative models and model-based reinforcement learning—offers a

productive bi-directional exchange. AI systems implementing active inference principles have demonstrated human-like exploration and learning in complex environments (Sajid et al., 2021), while biological insights from PP/AI have informed the design of more neurobiologically plausible artificial neural networks.

8. CONCLUSIONS

The predictive processing and active inference framework represents one of the most significant theoretical advances in cognitive science of the past quarter century. By grounding perception, cognition, action, emotion, and learning within a single computational principle—the minimization of variational free energy—the framework offers a degree of theoretical unification that no previous approach has achieved. Its neurobiological instantiation in hierarchical predictive coding provides testable mechanistic predictions that have, to a substantial degree, been confirmed by electrophysiological and neuroimaging evidence. Its clinical applications in computational psychiatry have generated novel, mechanistically grounded accounts of major mental health conditions and opened promising avenues for precision medicine.

Nonetheless, the framework faces significant challenges. Questions about falsifiability, the status of representation, and the relationship to alternative theories remain unresolved. The gap between computational-level descriptions and mechanistic-level explanations is a persistent tension that requires sustained empirical and theoretical effort to bridge. The framework's ambitions are perhaps most clearly stated in Friston's (2010) assertion that the FEP constitutes a 'variational free energy principle' that might unify biology, cognitive science, and physics—an ambition that remains as inspiring as it is contested.

In sum, predictive processing and active inference stand as a generative and transformative framework whose full scientific potential is yet to be realized. We anticipate that the coming decade will bring both important empirical tests of the framework's core predictions and productive dialogue between the PP/AI community and neighboring fields, ultimately producing a richer and more unified science of mind and brain.

AUTHOR NOTE

The authorship of this manuscript (Taruna Ikrar, Wachyudi Muchsin, Alfi Sophian) is identical to that of a separately submitted review on quantum consciousness by the same research group. These are genuinely distinct papers addressing different theoretical frameworks

in cognitive neuroscience and were developed independently. The authors confirm that the present manuscript represents wholly original work and has not been submitted elsewhere. Correspondence regarding this paper should be addressed to Alfi Sophian (indicated by asterisk in the author list).

References

Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. *Front Psychiatry*. 2013;4:47. doi:10.3389/fpsyt.2013.00047

Apps MAJ, Tsakiris M. The free-energy self: A predictive coding account of self-recognition. *Neurosci Biobehav Rev*. 2014;41:85-97. doi:10.1016/j.neubiorev.2013.01.029

Baars BJ. *A cognitive theory of consciousness*. Cambridge University Press; 1988.

Barlow HB. Possible principles underlying the transformation of sensory messages. In: Rosenblith WA, ed. *Sensory communication*. MIT Press; 1961:217-234.

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. *Neuron*. 2012;76(4):695-711. doi:10.1016/j.neuron.2012.10.038

Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*. 2015;85(2):390-401. doi:10.1016/j.neuron.2014.12.018

Botvinick M, Cohen J. Rubber hands 'feel' touch that eyes see. *Nature*. 1998;391(6669):756. doi:10.1038/35784

Clark A. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press; 2016.

Clark A. Consciousness as generative entanglement. *J Philos*. 2019;116(12):645-662. doi:10.5840/jphil20191161241

Colombo M. Why build a virtual brain? Large-scale neural simulations as test-bed for artificial computing systems. *Brain Cogn*. 2017;112:86-94.

Colombo M, Seriès P. Bayes in the brain—on Bayesian modelling in neuroscience. *Br J Philos Sci*. 2012;63(3):697-723. doi:10.1093/bjps/axr043

Corlett PR, Horga G, Fletcher PC, Alderson-Day B, Friston K, Powers AR. Hallucinations and strong priors. *Trends Cogn Sci*. 2019;23(2):114-127. doi:10.1016/j.tics.2018.12.001

Dayan P, Huys QJM. Serotonin in affective control. *Annu Rev Neurosci*. 2009;32:95-126. doi:10.1146/annurev.neuro.051508.135607

Dehaene S, Changeux JP, Naccache L. The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In: Dehaene S, Christen Y, eds. *Characterizing consciousness: From cognition to the clinic?*. Springer; 2011:55-84.

Egner T, Monti JM, Summerfield C. Expectation and surprise determine neural population responses in the ventral visual stream. *J Neurosci*. 2010;30(49):16601-16608. doi:10.1523/JNEUROSCI.2770-10.2010

Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*. 2002;415(6870):429-433. doi:10.1038/415429a

Feldman H, Friston KJ. Attention, uncertainty, and free-energy. *Front Hum Neurosci*. 2010;4:215. doi:10.3389/fnhum.2010.00215

Friston KJ. Hierarchical models in the brain. *PLoS Comput Biol*. 2008;4(11):e1000211. doi:10.1371/journal.pcbi.1000211

Friston KJ. The free-energy principle: A unified brain theory? *Nat Rev Neurosci*. 2010;11(2):127-138. doi:10.1038/nrn2787

Friston KJ, Daunizeau J, Kilner J, Kiebel SJ. Action and behavior: A free-energy formulation. *Biol Cybern*. 2010;102(3):227-260. doi:10.1007/s00422-010-0364-z

Friston KJ, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active inference: A process theory. *Neural Comput*. 2017;29(1):1-49. doi:10.1162/NECO_a_00912

Friston KJ, Kilner J, Harrison L. A free energy principle for the brain. *J Physiol Paris*. 2006;100(1-3):70-87. doi:10.1016/j.jphysparis.2006.10.001

Friston KJ, Shiner T, FitzGerald T, Galea JM, Adams R, Brown H, Dolan RJ, Moran R, Stephan KE, Bestmann S. Dopamine, affordance and active inference. *PLoS Comput Biol*. 2012;8(1):e1002327. doi:10.1371/journal.pcbi.1002327

Friston KJ, Wiese W, Hobson JA. Sentience and the free energy principle. *Phys Life Rev*. 2021;36:48-54. doi:10.1016/j.plrev.2020.12.002

Frith CD, Frith U. Mechanisms of social cognition. *Annu Rev Psychol*. 2012;63:287-313. doi:10.1146/annurev-psych-120710-100449

Garfinkel SN, Seth AK, Barrett AB, Suzuki K, Critchley HD. Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biol Psychol*. 2015;104:65-74. doi:10.1016/j.biopsycho.2014.11.004

Garrido MI, Kilner JM, Kiebel SJ, Friston KJ. The mismatch negativity: A review of underlying mechanisms. *Clin Neurophysiol*. 2009;120(3):453-463. doi:10.1016/j.clinph.2008.11.029

Helmholtz H von. *Handbuch der physiologischen Optik* [Handbook of physiological optics]. Voss; 1867.

Hohwy J. *The predictive mind*. Oxford University Press; 2013.

Hohwy J. New directions in predictive processing. *Mind Lang*. 2020;35(2):209-223. doi:10.1111/mila.12281

Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: An epistemological review. *Cognition*. 2008;108(3):687-701. doi:10.1016/j.cognition.2008.05.010

Huys QJM, Daw ND, Dayan P. Depression: A decision-theoretic analysis. *Annu Rev Neurosci*. 2015;38:1-23. doi:10.1146/annurev-neuro-071714-033928

Hutto DD, Myin E. *Radicalizing enactivism: Basic minds without content*. MIT Press; 2013.

Kant I. *Critique of pure reason*. (Guyer P, Wood AW, Trans.). Cambridge University Press; 1998. (Original work published 1781)

Kapur S. Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry*. 2003;160(1):13-23. doi:10.1176/appi.ajp.160.1.13

Keller GB, Mrsic-Flogel TD. Predictive processing: A canonical cortical computation. *Neuron*. 2018;100(2):424-435. doi:10.1016/j.neuron.2018.10.003

Kilner JM, Friston KJ, Frith CD. Predictive coding: An account of the mirror neuron system. *Cogn Process*. 2007;8(3):159-166. doi:10.1007/s10339-007-0170-2

Knill DC, Pouget A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci*. 2004;27(12):712-719. doi:10.1016/j.tins.2004.10.007

Kok P, Jehee JFM, de Lange FP. Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*. 2012;75(2):265-270. doi:10.1016/j.neuron.2012.04.034

Körding KP, Wolpert DM. Bayesian integration in sensorimotor learning. *Nature*. 2004;427(6971):244-247. doi:10.1038/nature02169

Kutas M, Federmeier KD. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annu Rev Psychol*. 2011;62:621-647. doi:10.1146/annurev.psych.093008.131123

Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. *Trends Cogn Sci*. 2012;16(1):72-80. doi:10.1016/j.tics.2011.11.018

Moutoussis M, Trujillo-Barreto NJ, El-Deredy W, Dolan RJ, Friston KJ. A formal model of interpersonal inference. *Front Hum Neurosci*. 2014;8:160. doi:10.3389/fnhum.2014.00160

Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I. 'Primitive intelligence' in the auditory cortex. *Trends Neurosci*. 2001;24(5):283-288. doi:10.1016/S0166-2236(00)01790-2

Parr T, Friston KJ. Generalised free energy and active inference. *Biol Cybern*. 2019;113(5-6):495-513. doi:10.1007/s00422-019-00805-w

Paulus MP, Stein MB. An insular view of anxiety. *Biol Psychiatry*. 2006;60(4):383-387. doi:10.1016/j.biopsych.2006.03.042

Pellicano E, Burr D. When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends Cogn Sci*. 2012;16(10):504-510. doi:10.1016/j.tics.2012.08.009

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*. 2006;442(7106):1042-1045. doi:10.1038/nature05051

Pickering MJ, Garrod S. An integrated theory of language production and comprehension. *Behav Brain Sci*. 2013;36(4):329-347. doi:10.1017/S0140525X12001495

Rao RPN, Ballard DH. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. 1999;2(1):79-87. doi:10.1038/4580

Sajid N, Ball PJ, Parr T, Friston KJ. Active inference: Demystified and compared. *Neural Comput*. 2021;33(3):674-712. doi:10.1162/neco_a_01357

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. *Science*. 1997;275(5306):1593-1599. doi:10.1126/science.275.5306.1593

Seth AK, Friston KJ. Active interoceptive inference and the emotional brain. *Philos Trans R Soc B*. 2016;371(1708):20160007. doi:10.1098/rstb.2016.0007

Stein MB, Simmons AN, Feinstein JS, Paulus MP. Increased amygdala and insula activation during emotion processing in anxiety-prone subjects. *Am J Psychiatry*. 2011;164(2):318-327. doi:10.1176/ajp.2007.164.2.318

Stephan KE, Binder EB, Breakspear M, Dayan P, Johnstone EC, Meyer-Lindenberg A, Schnyder U, Wang XJ, Bach DR, Fletcher PC, Friston KJ, Ganesh G, Garber H, Giurfa M, Iglesias S, Kasper S, Löffler-Stastka H, Murray RJ. Charting the landscape of priority problems in psychiatry, part 1: Classification and diagnosis. *Lancet Psychiatry*. 2016;3(1):77-83. doi:10.1016/S2215-0366(15)00465-2

Summerfield C, de Lange FP. Expectation in perceptual decision making: Neural and computational mechanisms. *Nat Rev Neurosci*. 2014;15(11):745-756. doi:10.1038/nrn3838

van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L, Wagemans J. Precise minds in uncertain worlds: Predictive coding in autism. *Psychol Rev*. 2014;121(4):649-675. doi:10.1037/a0037665

van Kerkoerle T, Self MW, Dagnino B, Gariel-Mathis MA, Poort J, van der Togt C, Roelfsema PR. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc Natl Acad Sci USA*. 2014;111(40):14332-14341. doi:10.1073/pnas.1402773111

Vossel S, Bauer M, Mathys C, Adams RA, Dolan RJ, Friston KJ, Stephan KE. Cholinergic stimulation enhances Bayesian belief updating in the deployment of spatial attention. *J Neurosci*. 2014;34(47):15735-15742. doi:10.1523/JNEUROSCI.0091-14.2014

Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. *Nat Neurosci*. 2002;5(6):598-604. doi:10.1038/nn858

Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. *Neuron*. 2005;46(4):681-692. doi:10.1016/j.neuron.2005.04.026