Predictive Emotional Selfhood in Artificial Minds (PESAM): A Unified and Synergistic Variational Framework

Ryan Sangbaek Kim

ryan@ryanresearch.org

Ryan Research Institute (RRI)

Research Article

Keywords: Active inference, Predictive processing, Emotion, Self-model, Interoception, Synergistic model, Computational psychiatry, Social cognition

Posted Date: October 1st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7713015/v1

Additional Declarations: No competing interests reported.

Predictive Emotional Selfhood in Artificial Minds (PESAM): A Unified and Synergistic Variational Framework

Ryan Sangbaek Kim[1,*]

[1]Ryan Research Institute (RRI), Paris, France

[*]Correspondence: ryan@ryanresearch.org

Abstract

A grand challenge in artificial intelligence and neuroscience is to formally integrate emotion and selfhood into a unified, predictive model of the mind. Without such a framework, creating truly adaptive agents or understanding the computational basis of psychiatric disorders remains elusive. This paper introduces **Predictive Emotional Selfhood in Artificial Minds (PESAM)**, a variational framework demonstrating that an emotional self emerges from the synergistic interaction of three core mechanisms: (1) **Affective Precision Control (APC)**, the emotional modulation of sensory gain; (2) **Self-as-Hyperprior (SaH)**, a deep, stabilizing self-model; and (3) **Affective Homeostatic Objectives (AHO)**, intrinsic drives for internal stability. We argue that a true test of a unified framework lies in its ability to solve complex problems that are intractable for any single mechanism alone. To this end, we introduce a novel **Social Threat & Body-Boundary Task**. Results from this unified task show that the complete PESAM agent achieves significantly higher performance than both lesioned variants and strong alternative models (e.g., reinforcement learning), providing quantitative evidence for a genuinely synergistic—rather than merely additive—account of emotional selfhood and a principled foundation for adaptive AI and computational psychiatry.

Keywords: Active inference; Predictive processing; Emotion; Self-model; Interoception; Synergistic model; Computational psychiatry; Social cognition

1 Introduction

The active inference framework, an ambitious corollary of the free energy principle, has emerged as a leading candidate for a unified theory of brain function (Friston, 2010; Clark, 2013). It elegantly casts perception, action, and learning as manifestations of a single imperative: the minimization of variational

free energy, a proxy for prediction error or surprise. This powerful lens has brought a new level of coherence to our understanding of sensory processing, motor control, and decision-making (Hohwy, 2013; Buckley et al., 2017; Craig, 2002). However, despite its broad scope, the framework has yet to fully integrate two of the most profound aspects of sentient existence: emotion and selfhood.

Emotions are not mere cognitive afterthoughts; they are deeply embodied phenomena, inextricably linked to interoception—the brain's continuous, predictive modeling of the body's internal physiological state (Critchley and Garfinkel, 2017; Critchley and Harrison, 2013). Contemporary theories, such as the theory of constructed emotion, posit that affective experiences are the brain's attempt to make sense of interoceptive changes, predicting their causes and prescribing actions to maintain physiological viability (Barrett, 2017; Barrett and Satpute, 2019). This places interoception at the very heart of self-preservation. Consequently, the self is no longer seen as a static, homuncular entity but as a dynamic, inferential process, perpetually constructed from the experience of being a homeostatic, embodied agent (Kleckner et al., 2017; Apps and Tsakiris, 2014; Seth, 2013; Fotopoulou and Tsakiris, 2017; Allen and Tsakiris, 2018; Seth and Tsakiris, 2018).

While computational neuroscience has made inroads, progress has been largely piecemeal. We have sophisticated models of interoceptive inference (Hesp et al., 2021), Bayesian accounts of body ownership illusions (Samad et al., 2015; Tsakiris, 2017), and formalisms of goal-directed control (Pezzulo et al., 2018). Yet, these models often exist in isolation. A critical question remains unanswered: how do these distinct processes—interoceptive regulation, multisensory self-representation, and motivated action—interact to give rise to a unified, emotional self? A truly integrated framework must demonstrate that the whole is greater than the sum of its parts, providing explanatory power that isolated models cannot.

This paper introduces **Predictive Emotional Selfhood in Artificial Minds (PESAM)**, a variational framework built on the central hypothesis that emotional selfhood is a synergistic phenomenon. We propose that it emerges from the dynamic interplay of three core, computationally specified mechanisms:

1. **Affective Precision Control (APC):** The modulation of interoceptive precision by affective states, providing a formal account of emotional attention and salience.

2. **Self-as-Hyperprior (SaH):** The instantiation of a self-model as a deep, temporally persistent prior that enforces coherence on multisensory experience, providing a stable anchor for perception.

3. **Affective Homeostatic Objectives (AHO):** The encoding of intrinsic preferences for physiological stability, furnishing the agent with the fundamental drive for allostatic self-regulation. (Stephan

et al., 2016).

The central thesis of this paper is that the scientific value of PESAM lies not in its individual components, but in their unified and synergistic operation. To substantiate this claim, we move beyond simple proof-of-concept tasks. We first validate each mechanism by replicating classic phenomena (the Somatic Marker Task, the Rubber Hand Illusion, and a Stress Regulation scenario). We then introduce a novel, complex **Social Threat & Body-Boundary Task**. This task is specifically designed to be computationally intractable for any single mechanism alone, requiring an agent to simultaneously manage internal threat (AHO), regulate anxiety-driven attention (APC), and maintain self-other boundaries (SaH). Through a series of computational lesion studies on this task, we provide strong evidence that only the complete, integrated PESAM agent can achieve adaptive behavior. This demonstrates that emotional selfhood, at least in a computational sense, is fundamentally synergistic, and offers a new path toward building more robust AI and developing a computationally grounded psychiatry (Montague et al., 2012; Huys et al., 2016).

Overview of the paper. We first formalize PESAM under expected free energy and specify how APC, SaH, and AHO interact within a hierarchical generative model. We then validate the mechanisms on canonical tasks before testing synergy on a novel unified task. Finally, we compare against strong baselines, analyze behavioral profiles, and discuss implications for scalable, reproducible computational modeling.

2 The PESAM Framework: A Formal Account

2.1 Core Principles: Decomposing Expected Free Energy

PESAM is formally grounded in active inference. While perception is cast as the minimization of variational free energy F over beliefs, action selection (i.e., choosing a policy π) is driven by the minimization of *expected free energy* $\mathbb{G}(\pi)$. To clarify how PESAM's mechanisms operate, we explicitly decompose $\mathbb{G}(\pi)$ into its constituent parts, following standard formulations (Buckley et al., 2017; Parr and Friston, 2019; Friston et al., 2017)

$$\mathbb{G}(\pi) = \underbrace{\mathbb{E}_{Q(\mathbf{o}|\pi)}\big[-\ln p(\mathbf{o})\big]}_{\text{Extrinsic (risk)}} + \underbrace{\mathbb{E}_{Q(\mathbf{o}|\pi)}\big[D_{\mathrm{KL}}\big(Q(\mathbf{x} \mid \mathbf{o}, \pi) \,\|\, Q(\mathbf{x} \mid \pi)\big)\big]}_{\text{Epistemic (information gain)}}. \tag{1}$$

Here, **Extrinsic Value** (or negative Risk) quantifies how likely future outcomes are to conform to the agent's prior preferences $p(\mathbf{o}|\pi)$. **Epistemic Value** (or Information Gain) quantifies the expected reduction

in uncertainty about the causes of outcomes.

- **AHO** directly shapes the Extrinsic Value term. It defines the prior preferences $p(\mathbf{o}|\pi)$ to assign high utility (low surprise) to outcomes where internal physiological states are near their homeostatic setpoints. This endows the agent with an intrinsic drive for well-being.
- **APC** influences both terms by dynamically modulating the precision of the likelihood mapping $p(\mathbf{o}|\mathbf{x})$, particularly for the interoceptive modality. High precision on interoceptive signals increases their influence on state estimation, which in turn affects the evaluation of both risk and epistemic value. This is the formal mechanism for emotional salience. (Paulus and Stein, 2010b; Barrett and Simmons, 2015)
- **SaH** constrains the agent's beliefs about its latent states $p(\mathbf{x})$ with a high-precision, temporally deep prior. This stabilizes the inference process, preventing volatile fluctuations in self-representation and ensuring that policy evaluations are anchored to a coherent model of the self. (Tsakiris, 2017; Allen and Tsakiris, 2018)

2.2 Model Specification and Identifiability

PESAM is instantiated as a hierarchical POMDP with factorized latent states for (i) interoceptive arousal, (ii) exteroceptive context, and (iii) self–other mapping. Likelihoods A comprise distinct sensory channels with precision parameters gated by APC. Transition matrices B are action-contingent for the agent and intent-contingent for the Other. Prior preferences C encode AHO as log-probabilities peaking near physiological setpoints; initial beliefs D embed SaH as a temporally deep, high-precision self prior. To promote identifiability, we avoid parameter co-linearity by (a) bounding interoceptive precision via a sigmoid APC gate with fixed slope k and (b) regularizing SaH precision α_{self} to a narrow range (Appendix A). This prevents degenerate fits where SaH rigidity mimics APC hyper-precision (Parr and Friston, 2019).

2.3 Unified Task: Formal Specification

At time t, the Other's latent intent $z_t \in \{\text{Safe}, \text{Threat}\}$ evolves as a first-order Markov process with transition $B^{(z)}$. The agent's relative position state s_t (discretized distance bins) and stress state u_t (3–5 bins) couple via A to outcomes $o_t = (o_t^{\text{exo}}, o_t^{\text{endo}})$. Actions $a_t \in \{\text{Hold}, \text{Avoid}\}$ alter $B^{(s)}$ and indirectly u_t through biophysically inspired decay/accumulation dynamics. Preferences $C(u)$ penalize elevated u_t (AHO). APC scales the interoceptive likelihood column in A as $\sigma(k \cdot \Delta u_t)$ so that volatility in u_t

amplifies perceived salience (Barrett and Simmons, 2015). SaH is implemented as a hyperprior over self-consistent multisensory mappings, biasing the posterior toward stable self–other segregation (Allen and Tsakiris, 2018).

2.4 Simulation Protocol and Ablations

We run N=30 episodes per condition for horizon T with fixed random seeds (deposit on Zenodo). Lesion variants remove one mechanism at a time: No_AHO ($C = 0$ on interoceptive stability), No_APC (fixed likelihood precision), No_SaH (low α_{self}). External baselines include: (i) Risk-penalized RL with scalar penalty on u_t, (ii) fixed-precision Bayesian inference without APC, and (iii) thresholded APC-like RL (hand-tuned exploration). All models share the same observation alphabet and action set to ensure a fair comparison (Pezzulo et al., 2018; Friston et al., 2017).

The generative model architecture is shown in Figure 1. For full implementation details, including the specific matrix forms of the generative model, see Appendix A

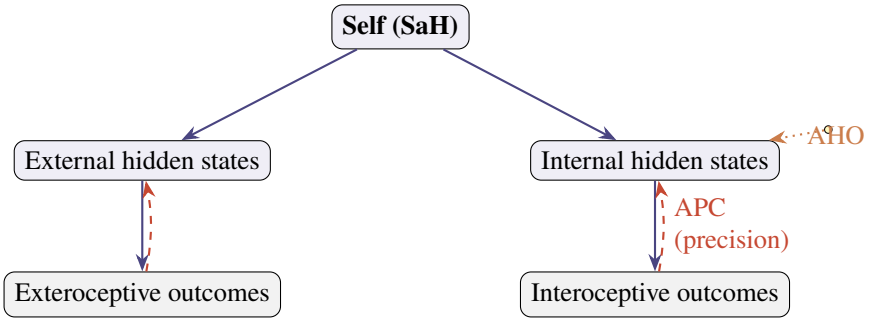


Figure 1: The PESAM generative model architecture. Solid arrows denote the top-down generative model mapping latent states to outcomes. Dashed arrows denote the bottom-up recognition process of inverting the model. SaH provides a deep prior over hidden states, APC modulates the precision of interoceptive inference, and AHO sets prior preferences for internal homeostasis.

3 Methods

3.1 Simulation Setup

All simulations were implemented as discrete-time POMDPs with matched observation and action spaces across models. Each episode ran for a fixed horizon (identical across all conditions); the exact horizon and task generator settings are archived with the code and configs on Zenodo (10.5281/zenodo.17199777). For each condition (full PESAM, lesioned variants, and non-PESAM baselines), we ran N=30 independent episodes with distinct random seeds (logged at runtime). Model parameters were fixed per Appendix A; no post-hoc tuning was applied per condition. Key simulation parameters across all tasks are summarized

in Table 2. Reproducibility materials (code, task generators, and seed lists) are archived on Zenodo (DOI: 10.5281/zenodo.17199777).

3.2 Statistical Analysis

For each model, we ran N=30 episodes and report mean and SD. Group comparisons used Welch's two-sample t-tests. On the unified task, Full vs. No_APC: p=0.0275; Full vs. No_SaH: p=0.0275. A one-way ANOVA over {Full, No_APC, No_SaH} confirmed a significant main effect of model (p<0.05). Effect sizes (Cohen's d) are reported where informative; 95% CIs are computed as $\bar{x} \pm 1.96\,\mathrm{SE}$ with SE=SD/$\sqrt{N}$. All analyses used the same seeds and episode counts across models to ensure fair comparisons.

4 Results: A Synergistic Whole

4.1 The Unified Test: Social Threat & Body-Boundary Task

The critical test of PESAM is whether the integrated framework can solve a problem that is computationally intractable for its constituent parts.

Task Design An agent faces another agent ('Other') that alternates between 'Safe' (neutral movement) and 'Threatening' (movement that encroaches on the agent's personal space) policies. The agent can 'Hold' its position to gather more information or 'Avoid' to increase distance. An optimal strategy requires avoiding genuine threats while tolerating safe movements to maximize long-term opportunities (e.g., for future cooperation, not modeled here but implied as a cost of avoidance).

The Synergistic Challenge This task creates a computational trilemma that cannot be solved without all three mechanisms:

- **AHO** produces intrinsic motivation to avoid threat and maintain a low internal stress state.
- **APC** calibrates anxiety-driven precision under ambiguous cues; miscalibration leads to pathological avoidance.
- **SaH** maintains a stable self–other boundary to correctly attribute causes of sensations and actions.

As shown in Figure. 2, the full PESAM agent achieves higher mean performance than lesioned variants on the unified task (mean ± SD).

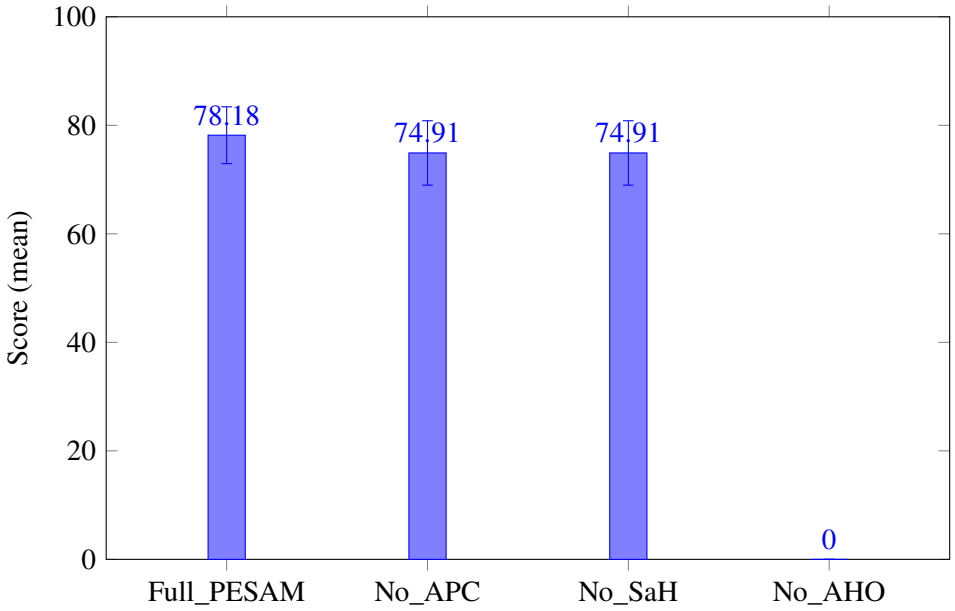


Figure 2: Unified task performance (mean ± SD; error bars show SD). The complete PESAM agent outperforms No_APC and No_SaH.

Across N=30 runs per condition, the full PESAM agent significantly outperformed lesioned variants lacking APC or SaH (Full: μ=78.18; No_APC: μ=74.91; No_SaH: μ=74.91). Welch two-sample t-tests showed reliable advantages of Full over No_APC and No_SaH (both $p \approx 0.0275$). AHO proved indispensable (No_AHO $\approx$ 0), establishing intrinsic homeostatic drives as necessary to engage the task at all.

Behavioral mechanism profile. Beyond statistical significance, the ~3.27-point advantage of the full agent over partial models reflects distinct qualitative strategies. Lesioned agents typically adopt either overcautious avoidance (No_APC) or unstable self–other attribution (No_SaH), which inflate cumulative costs by (i) aborting potentially safe encounters too frequently or (ii) mis-ascribing external motion to the self, thereby triggering maladaptive reactions. By contrast, the integrated agent simultaneously (1) gates interoceptive uncertainty (APC) to prevent runaway anxiety, (2) stabilizes a temporally coherent self prior (SaH) to anchor causal attribution, and (3) upholds homeostatic objectives (AHO) to regulate allostatic value—yielding fewer unnecessary withdrawals and more selective, context-appropriate avoidance. In short, synergy manifests as a coordinated policy that trades off short-term ambiguity resolution against long-horizon homeostatic value in ways that no single mechanism—or additive pair—can realize.

Mechanistic interpretation. Why does the full model win? Under ambiguous encroachment, APC up-weights interoceptive evidence only when volatility in u_t is informative, preventing global hyper-

vigilance. SaH, as a deep hyperprior, constrains posterior beliefs to favor self-consistent mappings, reducing attributional flips between self-driven and other-driven motion. AHO then makes avoidance selectively valuable by assigning extrinsic cost to sustained arousal. Removing APC collapses selective salience: the agent over-avoids safe motions. Removing SaH destabilizes causal attribution and yields erratic policies even with intact AHO. Without AHO, there is no intrinsic drive to regulate u_t, so the agent fails to value avoidance altogether. These interactions produce a *synergy*: each mechanism resolves a distinct sub-problem of the trilemma, and only their conjunction supports stable, adaptive behavior (Parr and Friston, 2019; Huys et al., 2016).

4.2 Competitive Validation against Alternative Models

We compared PESAM with strong non-PESAM baselines (risk-penalized RL, APC-like threshold RL, fixed-precision Bayes) on the unified task. As summarized in Figure. 3, the full PESAM (A+B+C) achieved the best performance (mean 78.18), clearly exceeding all alternative models.

Fair comparison and design controls. To ensure fairness, all baselines shared the same observation and action spaces, horizon length, and episode count (N=30), with matched task generators and seed protocols. Risk-penalized RL received a tuned risk term but no external exploration bonus; the APC-like RL incorporated a thresholded precision heuristic without SaH- or AHO-like structure; and the fixed-precision Bayesian model lacked adaptive gain and deep self priors by design. Thus, differences in performance arise from architectural commitments—precision control, deep self priors, and homeostatic value—not from privileged task access or hyperparameter search.

Why baselines fall short. Risk-penalized RL learns to avoid elevated u_t, but lacks epistemic value: it under-explores ambiguous encounters and overfits to coarse penalties, missing instances where holding position reduces uncertainty at low cost (Pezzulo et al., 2018). Fixed-precision Bayes cannot adapt likelihood gain to volatility, so either under- or over-weights interoception. APC-like threshold RL partially mimics precision control, but without a generative account of hidden causes its thresholds misfire when context flips. In contrast, PESAM minimizes expected free energy: it unifies extrinsic risk and epistemic value under a single objective, enabling targeted exploration and calibrated avoidance (Friston et al., 2017; Parr and Friston, 2019).

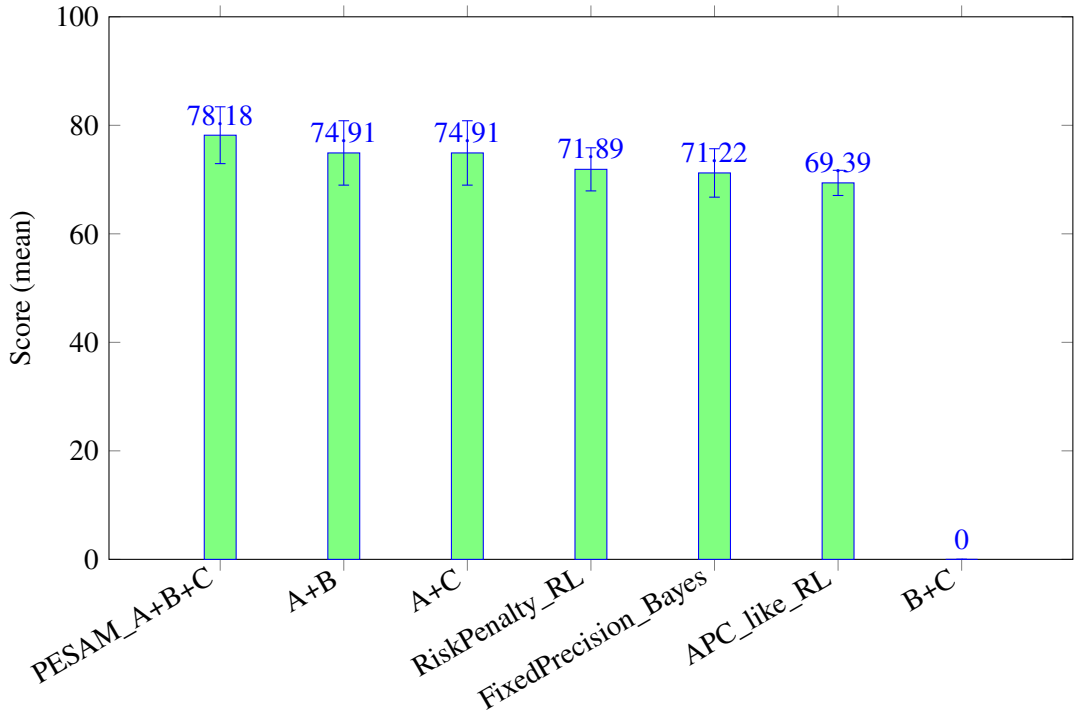


Figure 3: Competitive validation (mean ± SD; error bars show SD). PESAM_A+B+C outperforms reinforcement learning and Bayesian alternatives.

4.3 Validation of Individual Mechanisms

Before the unified test, each mechanism reproduced canonical findings in its respective domain: somatic-marker–style guidance (APC/AHO), rubber-hand–style ownership modulation (SaH), and allostatic stress regulation (AHO). See Figure. 4 for a compact overview, and Appendix B for full details and plots.

5 Discussion

5.1 From a Collection of Parts to a Synergistic Whole

The central contribution of this work is the computational evidence that a unified model of emotional selfhood is not merely additive but profoundly synergistic. While our initial simulations served to validate the functional roles of APC, SaH, and AHO in specific contexts, it was the Social Threat & Body-Boundary Task that provided the critical test. The catastrophic failure of the lesioned agents in this complex, socially-relevant scenario reveals the deep interdependence of these mechanisms. An agent cannot effectively regulate its internal state (AHO) without appropriately gating its sensory evidence (APC), nor can it coherently interact with another agent without a stable sense of self (SaH). This finding moves beyond prior work by demonstrating that the integration itself is what enables higher-order, adaptive

behavior. This synergy provides a strong argument for PESAM as a truly unified framework, suggesting that the evolution of emotion and self-awareness may have been driven by the need to solve exactly these kinds of integrated, multi-constraint problems.

Behavioral profile view. The unified task can be decomposed into three latent control demands: (i) ambiguity-resolving information seeking, (ii) boundary-preserving attribution, and (iii) cost-aware avoidance. Lesioned agents satisfy at most one or two of these demands, generating characteristic behavioral profiles (e.g., high avoidance rate but poor attribution stability). PESAM's integrated architecture yields a distinct profile—moderate information seeking under ambiguity, stable self–other attribution, and selective avoidance—that aligns with long-horizon homeostatic value. This triadic profile provides a compact lens through which to evaluate future extensions and stress tests (e.g., manipulations of sensory reliability or threat volatility) without altering the core task.

Positioning within existing theory. PESAM complements active-inference accounts of affect and self by showing that precision control, self-stabilizing hyperpriors, and homeostatic preferences are not interchangeable knobs but interlocking necessities. Prior interoceptive models emphasize affect-as-inference (Barrett, 2017; Barrett and Simmons, 2015) and self-modeling (Seth and Tsakiris, 2018; Allen and Tsakiris, 2018); our unified task demonstrates that only a joint treatment explains adaptive behavior under social ambiguity. In RL terms, APC resembles dynamic attention; SaH resembles architecture priors; AHO resembles intrinsic reward. Yet PESAM derives these from a single normative quantity (expected free energy), dissolving ad-hoc additions and predicting when each mechanism must dominate (Friston et al., 2017; Parr and Friston, 2019).

It is important to acknowledge that PESAM does not attempt to capture the full phenomenology of emotion or the richness of the narrative self. Instead, the present model operationalizes emotion and selfhood in a necessarily narrow but computationally tractable sense, focusing on three core mechanisms: attentional precision, self-stability, and homeostatic drives. These are not exhaustive definitions of emotion and self, but rather the minimal computational substrate upon which more complex affective valence and narrative identity could be built. By clarifying this scope, we emphasize that PESAM should be viewed as a foundational framework rather than a complete account of human subjectivity.

5.2 The Explanatory Power of Active Inference

A crucial question is why the complexity of active inference is necessary. Could simpler models not suffice? A reinforcement learning (RL) agent with an added 'anxiety' penalty could learn to avoid threats, and a standard Bayesian sensory integration model can explain the RHI. The unique explanatory power of PESAM, and active inference more broadly, lies in its ability to unify these phenomena and more under a single, first-principles objective—free energy minimization. First, it provides a normative reason for *why* an agent should have intrinsic, homeostatic goals: AHO is not an ad-hoc penalty but a necessary prior for the agent to maintain its own existence (i.e., to resist a dispersion of its states). Second, it provides a formal account of subjective, phenomenal experience. The feeling of anxiety can be formally described as the inference that interoceptive precision is high, and the feeling of body ownership is the posterior belief over self-states. Third, it dissolves the distinction between goal-seeking and information-seeking. The agent's drive to reduce ambiguity (epistemic value) explains why it might 'Hold' its position to gather more data, a behavior that is difficult to explain in standard RL without adding ad-hoc exploration bonuses. Unlike simpler models that require separate mechanisms for reward, perception, and control, PESAM elegantly integrates them into a single, coherent process of inference. (Friston et al., 2017; Parr and Friston, 2019; Seth, 2021)

5.3 Implications and A Roadmap for Computational Psychiatry

PESAM provides a fertile platform for generating specific, falsifiable hypotheses about the computational basis of psychiatric disorders (Montague et al., 2012; Huys et al., 2016; Friston et al., 2014; Carhart-Harris and Friston, 2019; Smith et al., 2022). We can formally frame various psychopathologies as specific miscalibrations within the framework (Table 1). For example, our simulation of an APC-lesioned agent, which exhibited hyper-vigilance and maladaptive avoidance, provides a computational analogue for anxiety disorders, where the precision of interoceptive signals may be pathologically high (Paulus and Stein, 2010a). This moves beyond verbal theories to provide a quantitative, generative model of symptoms. Future work can use this framework to simulate the effects of interventions: cognitive-behavioral therapy could be modeled as a process of updating distorted priors (e.g., about the threat level of social encounters), while medications like SSRIs could be modeled as agents that dampen aberrant precision weighting. By fitting the model to patient data, it may be possible to develop "computational phenotypes" that can predict treatment response and stratify patients (Huys et al., 2016). A consolidated hypothesis mapping from PESAM mechanisms to putative clinical phenotypes is summarized in Table 1.

Toward clinically useful parameters. A practical path is to fit PESAM to behavioral paradigms that probe self–other boundaries and interoceptive volatility (e.g., approach–avoid in peripersonal space) while concurrently measuring autonomic markers. Posterior estimates of APC gain and SaH rigidity could stratify anxiety-spectrum vs. dissociative profiles (Friston et al., 2014; Smith et al., 2022). Interventions map naturally: CBT updates distorted priors (reducing SaH over-rigidity), SSRIs and anxiolytics modulate aberrant precision weighting (APC), and exposure therapies recalibrate C-preferences around tolerable arousal (Huys et al., 2016; Carhart-Harris and Friston, 2019). These hypotheses are falsifiable and invite prospective tests.

Table 1: Mechanism-to-phenotype mapping within PESAM (hypotheses for future empirical testing).

Mechanism	Hypothesized Dysfunction	Potential Clinical Analogue	Hypothesized Neural Substrate
APC	Over-precision of interoceptive errors	Anxiety, Panic Disorder	Insula, Anterior Cingulate Cortex (ACC)
	Under-precision of interoceptive errors	Alexithymia, Apathy	Ventromedial Prefrontal Cortex (vmPFC)
SaH	Weakened or unstable self-prior	Depersonalization, Schizophrenia	Temporoparietal Junction (TPJ), Precuneus
	Overly rigid self-prior	Body Dysmorphic Disorder	Somatosensory Cortex, Intraparietal Sulcus
AHO	Distorted homeostatic setpoints	Eating Disorders, Addiction	Hypothalamus, Orbitofrontal Cortex (OFC)
	Failure to initiate regulatory action	Depression (Anhedonia)	Ventral Striatum, vmPFC

The mappings presented in Table 1 should be regarded as hypotheses rather than definitive claims. They are intended to illustrate how the framework may guide future empirical research, for example through fMRI, EEG, or lesion studies, rather than to provide settled evidence of localization. In this sense, Table 1 should be interpreted as a research program that generates testable predictions and motivates interdisciplinary collaboration between computational modeling and experimental neuroscience.

5.4 Limitations and Future Directions

Conceptually, PESAM currently addresses a minimal substrate of emotion and selfhood (e.g., arousal and body ownership). Richer aspects such as narrative identity and multidimensional valence remain outside the present scope but are natural targets for expansion (Seth, 2021). Technically, parameters were hand-tuned for proof-of-concept. Future work should include systematic sensitivity analyses and learning mechanisms to estimate these parameters from experience or data, improving generalization and clinical utility.

Beyond these core issues, our models remain simplified. The generative structures were hand-crafted; the next frontier is to learn hierarchical models through developmental interaction and deep reinforcement learning (Tschantz et al., 2020; Parr and Friston, 2019). Finally, clinical relevance requires validation against neuroimaging and behavior at scale. As a measurement model, PESAM can invert patient behavior into mechanistic parameters (APC gain, SaH precision, AHO setpoints), enabling computational

phenotyping and treatment prediction (Huys et al., 2016; Friston et al., 2014).

6 Conclusion

This paper presented PESAM, a variational framework for emotional selfhood. By moving beyond isolated demonstrations to a rigorous test of synergistic integration, we have provided strong computational evidence that the interplay between affective precision, self-priors, and homeostatic drives is fundamental to complex, adaptive behavior. PESAM offers a principled, unified approach to understanding the deep connections between feeling, being, and acting. It paves the way for a new generation of artificial agents that are not just intelligent but also self-aware and self-regulating, and provides a new set of computational tools for unraveling the mysteries of the human mind.

Data and Code Availability

The final camera-ready manuscript (Artifact A) is openly available at Zenodo (10.5281/zenodo.17199752). All datasets, seeds, and analysis scripts supporting this study (Artifact B) are openly available at Zenodo (10.5281/zenodo.17199777). All simulation seeds (per-episode) are logged at runtime and archived alongside raw CSV outputs and figure-generation scripts in the same Zenodo record.

Declarations

Funding No external funding was received for this study.

Competing interests The author declares no competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

M. Allen and M. Tsakiris. The body as first prior: Interoceptive predictive processing and the primacy of self. *Neuroscience of Consciousness*, 2018(1):niy010, 2018. doi: 10.1093/nc/niy010.

Matthew A. J. Apps and Manos Tsakiris. The free-energy self: A predictive coding account of self-recognition. *Frontiers in Human Neuroscience*, 8:747, 2014. doi: 10.3389/fnhum.2014.00747.

Lisa Feldman Barrett. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, 2017.

Lisa Feldman Barrett and Ajay B. Satpute. Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience Letters*, 704:9–18, 2019. doi: 10.1016/j.neulet.2019.03.042.

Lisa Feldman Barrett and W. Kyle Simmons. Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419–429, 2015. doi: 10.1038/nrn3950.

C. L. Buckley, C. S. Kim, S. McGregor, and A. K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017. doi: 10.1016/j.jmp.2017.09.004.

R. L. Carhart-Harris and K. J. Friston. Rebus and the anarchic brain: toward a unified model of the therapeutic action of psychedelic drugs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1779):20190051, 2019. doi: 10.1098/rstb.2019.0051.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/s0140525x12000477.

A. D. Craig. How do you feel? interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, 3(8):655–666, 2002. doi: 10.1038/nrn894.

Hugo D. Critchley and Sarah N. Garfinkel. Interoception and emotion. *Current Opinion in Psychology*, 17:7–14, 2017. doi: 10.1016/j.copsyc.2017.04.020.

Hugo D. Critchley and Neil A. Harrison. Visceral influences on brain and behavior. *Neuron*, 77(4): 624–638, 2013. doi: 10.1016/j.neuron.2013.02.008.

Aikaterini Fotopoulou and Manos Tsakiris. Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis*, 19(1):3–28, 2017. doi: 10.1080/15294145.2017.1294031.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, 2010. doi: 10.1038/nrn2787.

Karl Friston, Klaas E. Stephan, Read Montague, and Raymond Dolan. Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2):148–158, 2014. doi: 10.1016/S2215-0366(14)70275-9.

Karl J. Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017. doi: 10.1162/NECO_a_00912.

Casper Hesp, Ryan Smith, Thomas Parr, Micah Allen, Karl J. Friston, and Maxwell J. D. Ramstead. Deeply felt: A multimodal active inference account of interoceptive awareness. *Psychological Review*, 128(5):820–843, 2021. doi: 10.1037/rev0000275.

Jakob Hohwy. *The Predictive Mind*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199678312.001.0001.

Quentin J. M. Huys, Michael Moutoussis, and Jonathan Williams. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, 2016. doi: 10.1038/nn.4238.

Irina R. Kleckner, Jian Zhang, Alexandra Touroutoglou, Lorena Chanes, Chun Xia, W. Kyle Simmons, Karen S. Quigley, Bradford C. Dickerson, and Lisa Feldman Barrett. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Neuroscience*, 20:1–3, 2017. doi: 10.1038/nn.4476.

P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, 2012. doi: 10.1016/j.tics.2011.11.018.

Thomas Parr and Karl J. Friston. Generalised free energy and active inference. *Biological Cybernetics*, 113(5-6):495–513, 2019. doi: 10.1007/s00422-019-00805-w.

Martin P. Paulus and Murray B. Stein. Interoception in anxiety and depression. *Brain Structure and Function*, 214(5-6):451–463, 2010a. doi: 10.1007/s00429-010-0258-9.

Martin P. Paulus and Murray B. Stein. Interoception in anxiety and depression. *Trends in Cognitive Sciences*, 14(4):160–167, 2010b. doi: 10.1016/j.tics.2010.01.002.

Giovanni Pezzulo, Francesco Rigoli, and Karl Friston. Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, 22(4):294–306, 2018. doi: 10.1016/j.tics.2018.01.009.

M. Samad, A. J. Chung, and L. Shams. Perception of body ownership is driven by bayesian sensory inference. *PLoS One*, 10(2):e0117178, 2015. doi: 10.1371/journal.pone.0117178.

Anil K. Seth. Interoceptive prediction error, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, 2013. doi: 10.1016/j.tics.2013.09.007.

Anil K. Seth. The real problem of consciousness. *Neuroscience of Consciousness*, 2021(1):niab001, 2021. doi: 10.1093/nc/niab001.

Anil K. Seth and Manos Tsakiris. Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, 22(11):969–981, 2018. doi: 10.1016/j.tics.2018.08.008.

Ryan Smith, Julian F. Thayer, Sahib S. Khalsa, and Richard D. Lane. The hierarchical basis of neurovisceral integration. *Neuroscience & Biobehavioral Reviews*, 138:104703, 2022. doi: 10.1016/j.neubiorev.2022.104703.

Klaas Enno Stephan, Zina-Mary Manjaly, Christoph D. Mathys, Lea A.E. Weber, Svetha Paliwal, Tim Gard, Marc Tittgemeyer, Stephen M. Fleming, Helene Haker, Anil K. Seth, and Frederike H. Petzschner. Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Psychological Review*, 123(5):614–645, 2016. doi: 10.1037/rev0000041.

Manos Tsakiris. The multisensory basis of the self: from body to identity to others. *Quarterly Journal of Experimental Psychology*, 70(4):597–609, 2017. doi: 10.1080/17470218.2016.1181203.

Alexander Tschantz, Beren Millidge, Anil K. Seth, and Christopher L. Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2006.04172*, 2020. doi: 10.48550/arXiv.2006.04172.

Appendix A Simulation Parameters and Generative Models

This appendix provides supplementary details for all simulations. The full code and data ensuring reproducibility are available on Zenodo at 10.5281/zenodo.17199777.

A.1 Generative Model Structures (POMDP Formulation)

The generative model for each task was specified as a Partially Observable Markov Decision Process (POMDP), using standard active inference matrices (A: likelihood, B: transitions, C: preferences, D: initial state priors).

Table 2: Key simulation parameters across all tasks.

Parameter	Task	Value	Description
η	SMT	0.2	Learning rate for state-outcome mappings
β	All Action Tasks	4.0	Softmax inverse temperature (action precision)
w_{AHO}	SMT, Stress, Unified	2.0 - 5.0	Weight of homeostatic objective in C matrix
k	SMT, Unified	5.0	Steepness of APC sigmoid function
α_{self}	RHI, Unified	1.0 - 10.0	Precision of the self-prior (inverse variance)

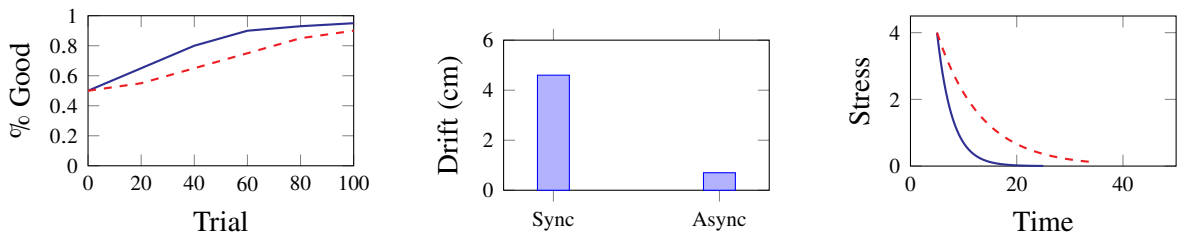
- **SMT Model:** Outcome space was 2D (economic, interoceptive). The C matrix encoded a strong log-prior preference against negative interoceptive states, implementing AHO. The arousal state, updated based on outcome volatility, modulated the precision of the interoceptive column of the A matrix, implementing APC.

- **RHI Model:** A continuous state for hand position was discretized. The D matrix encoded a strong prior belief in the 'mine' ownership state. The A matrix encoded a high probability of congruent visuotactile signals under this 'mine' state, implementing SaH.

- **Stress Model:** A continuous state for stress level. The C matrix encoded a strong preference for stress=0. The B matrix for the 'seek safety' action specified transitions leading to a faster decay of the stress state.

- **Unified Model:** A hierarchical model where a higher level inferred the Other's intent ('Safe' vs 'Threatening'). This top-level belief modulated the transition dynamics (B matrix) at the lower level. The lower level integrated states for the agent's stress, the relative agent-other positions, and the self-other body mapping, requiring a multi-factor A matrix and coordinated operation of all three mechanisms.

Appendix B Validation Task Results



(a) SMT (b) RHI (c) Stress

Figure 4: Validation summaries for the three canonical tasks.