Valence Computation as Higher-order Inference via Conceptual Self-Processing

Yuyue Jiang* **(yuyuejiang@ucsb.edu)**
Department of Psychological & Brain Sciences, University of California Santa Barbara
Santa Barbara, CA 93106 USA

Dezhi Luo* **(ihzedoul@umich.edu)**
Weinberg Institute for Cognitive Science, University of Michigan
Ann Arbor, MI 48109 USA

* Equal Contributions.

Abstract

Predictive processing models suggest that emotions arise from the hierarchical computation of prediction errors across signals. However, this framework alone cannot account for the subjective, evaluative quality of emotional experience. Here, we extend predictive processing by integrating higher-order theories of emotion, proposing that emotional valence emerges from value judgments grounded in conceptual self-processing over sensory and contextual representations. This framework offers a mechanistic account of how subjective emotional experience arises through inferences about the dynamic interplay between world- and self-models within a shared computational architecture.

Keywords: valence; self-prcessing; predictive processing; metacognition; subjective experience

Introduction

The subjective experience of valence—how something feels good or bad—is a defining feature of emotional experience, yet it is often underexplained in mechanistic models. Predictive processing (PP) offers a powerful framework, describing emotions as Bayesian inferences that minimize prediction errors (PEs) by integrating interoceptive and exteroceptive signals with prior expectations (Clark, 2013; Barrett, 2017). Emotions emerge when mismatches between predictions and sensory inputs signal the need for adaptive updating—for instance, unexpected social rejection may trigger sadness by violating social priors (Mobbs et al., 2019). While this model captures key aspects of emotional dynamics, it tends to formalize valence and arousal as mere functions of prediction error properties (Joffily & Coricelli, 2013), risking a conflation of emotional experience with generic state with heightened salience and/or surprise, and leaving the subjective, evaluative quality of emotion insufficiently accounted for.

On the other hand, higher-order theories of emotion offer a compelling account of emotional experience by emphasizing the role of conceptual self-processing (LeDoux & Brown, 2017). A central claim of these theories is that emotions arise when self-schemas are integrated into a conceptual representation of the scenario—encompassing both the stimulus and its context. This self-related layer of representation is seen as critical for emotional experience. As such, higher-order theories are well-suited to explaining the felt quality of valence, explicitly linking emotional experience to metacognitive awareness. However, in contrast to predictive processing, these theories offer limited mechanistic specificity. It remains unclear how self-related representations are instantiated in the brain or how they interact with lower-level perceptual and interoceptive processes.

Here, we extend existing models by integrating conceptual self-processing into the predictive processing framework to account for the subjective dimension of emotional experience. We propose that emotions arise when an individual not only predicts and integrates affective signals to form a coherent representation of the stimulus and its context, but also performs higher-order value judgments over this world-model, leveraging distilled conceptual information grounded in a dynamic self-model.

Neural Basis of Hierarchical Inference Underlying Emotions

We propose that emotional experience unfolds across three hierarchical stages of neural processing. The first two stages are consistent with standard predictive processing (PP) models, which frame emotions as emerging from the prediction errors across interoceptive and exteroceptive inputs and their conceptualization (Hohwy, 2020; Seth, 2013). First, subcortical systems integrate these sensory signals to generate rapid, survival-oriented affective responses (Barrett, 2017). These non-conscious states guide behavior by highlighting biologically salient stimuli and initiating autonomic reactions. Second, cortical networks incorporate contextual and semantic information to refine these affective predictions and reduce ambiguity (e.g., “Is this dangerous?”). At this level, emotional signals are reinterpreted through situational knowledge and conceptual associations, but still lack subjective, first-person significance.

The third stage, by contrast, extends beyond the standard PP framework and aligns with predictions from higher-order theories of emotion. We argue that subjective emotional experience—and the evaluation of valence—requires an additional layer of conceptual self-processing (LeDoux & Brown, 2017; LeDoux & Lau, 2020). This involves binding higher-order representations of a situation to a dynamic self-model comprising schematic knowledge of past experiences and associated propositional attitudes (Jiang & Luo, 2024). Crucially, this self-related inference goes beyond lower-order, reflexive representations and entails abstract evaluation of the situation’s relevance to the self. Valence, on this view, is not computed merely as a function of prediction error properties, but as a higher-order inference about self-relevance, shaped by internalized beliefs, goals, and values.

This integration is supported by a network centered on the ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex (ACC), and medial temporal lobe (MTL). Prior research has consistently implicated the vmPFC in memory-based value judgment and decision-making, particularly through its role in indexing schematic information from autobiographical memory (Hampton et al., 2006; Hebscher & Gilboa, 2016; Vaidya & Badre, 2020). More recent findings suggest that vmPFC specifically supports the evaluation of personal significance, rather than normative value, reinforcing its central role in self-referential processing (D’Argembeau, 2013; Kim & Johnson, 2015). This aligns with longstanding theories emphasizing the use of autobiographical memory schema in self-related evaluation and decision-making (Kihlstrom et al., 1988; Conway & Pleydell-Pearce, 2000; Prebble et al., 2013).

Notably, this third stage of self-processing does not require a separate mechanism outside the predictive framework, as some traditional higher-order theories imply. Rather, the in-

tegration of self-schematic information is consistent with the same higher-order inference mechanisms involved in implicit metacognition for perceptual monitoring (Dijkstra et al., 2022; Bein & Niv, 2025). Both processes rely on medial prefrontal-centered networks that perform dimensionality reduction and guide memory reactivation through interactions with posterior brain regions. The distinction between emotional valence and perceptual metacognition is thus content-selective, driven by a specialization for self-referential information in the vmPFC.

Formalization of Valence Computation Within the Higher-order PP Framework

We provide a high-level formalized account in which emotional experience arises from the interaction between two internal systems: the world-model and the self-model (Johnson-Laird, 1983; LeCun, 2022; Jiang & Luo, 2024). While unified by shared predictive processing mechanisms, these systems diverge in the content they access and the inferences they generate.

The world-model receives input in the form of an external stimulus, denoted S_x. The world-model starts by computing an intrinsic value based on current interoceptive and homeostatic states. This value reflects bottom-up bodily signals. To evaluate the meaning of this stimulus, the model performs a similarity-based search over a structured factual memory base off semantic knowledge and contextual representations. The value function returns the most similar knowledge and retrieves its associated information for further usage. For initial evaluation, the value associated with an stimulus is defined as an linear combination of the intrinsic value and the value learned from prior indexing (D'Argembeau, 2013). This operation thus supports an recursive evaluation based on conceptualization and generalization, allowing for the process novel inputs based on existing knowledge:

$$M_{S_x} \leftarrow \arg\max_{M_i \in Fact} Similarity(S_x, M_i)$$

$$Value(S_x) = Value(M_{S_x}) + IntrisicValue(S_x))$$

Once this associated value is computed, it is compared to a previously constructed value for the predicted stimulus, allowing the system to compute a prediction error:

$$PE_{world} = Value(S_x) - Value(Pred)(S_x)$$

The prediction error computed in the world-model then serves as a precision-weighting signal for higher-order inference within the self-model. First, the self-model maps S_x and its associated prediction error into a high-dimensional identity space. This transformation produces a vector M_{self}, representing the conceptual information associated with self-referential memory schema (LeDoux, 2020; Bein & Niv, 2025). This mapping operation relies on semantic transformation mechanisms, which we can denote as:

$$m_{self} = \Phi(S_x, PE_{world})$$

To evaluate the relevance of m_{self}, the self-model performs a second similarity-based memory retrieval—this time from the autobiographical memory system (Conway & Pleydell-Pearce, 2000). This system stores emotionally tagged, self-relevant episodes from past experiences. The retrieval yields two key outputs: a self-representation S_R and an associated valence V_{S_R}. The former reflects the version of the self most relevant to the current stimulus, while the latter encodes the valence of the past episode based on previous inferences.

$$[S_R, V_{S_R}] \leftarrow \arg\max_{M_i \in M_{Auto}} Similarity(m_{self}, M_i)$$

The autobiographical retrieval process is supported by the hippocampus and medial temporal lobe structures, which enable episodic simulation, as well as the ventromedial prefrontal cortex, which integrates memory with value and identity-related features.

The next step in the self-model involves computing a *self-prediction error*—that is, the mismatch between the current self-representation M_{self} and the self-state the model would have predicted in this situation, denoted $\hat{M}_{self}$. This mismatch is not an arithmetic difference but a semantic distance in identity space. This operation reflects psychological conflict or coherence: how closely the actual meaning of a situation aligns with one's expectation.

The final computation within the self-model is the emotional valence itself. This value results from integrating multiple components: the self-prediction error, the retrieved valence of the self-representation, and the interaction between self and world-model prediction errors. The formal expression for this integration is as follows:

$$Valence = \alpha \cdot PE_{self} \cdot SR + \beta \cdot V_{SR}$$

The first term captures how the accuracy of one's self-predictions contributes to affective tone. For instance, a close match between M_{self} and $\hat{M}_{self}$ may yield confidence or pride, while a discrepancy may induce guilt or shame. The second term ensures emotional continuity, allowing the system to ground present emotions in past affective states, reflecting a consistent emotion schema at individual level.

Conclusion

We propose a higher-order predictive processing account of emotion, in which valence computation is formalized as value judgment instantiated through higher-order inference via conceptual self-processing.

References

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

Bein, O., & Niv, Y. (2025). Schemas, reinforcement learning and the medial prefrontal cortex. *Nature Reviews Neuroscience*, 1–17.

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

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological review*, *107*(2), 261.

Dijkstra, N., Kok, P., & Fleming, S. M. (2022). Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. *Neuroscience & Biobehavioral Reviews*, *135*, 104557.

D'Argembeau, A. (2013). On the role of the ventromedial prefrontal cortex in self-processing: the valuation hypothesis. *Frontiers in human neuroscience*, *7*, 372.

Hampton, A. N., Bossaerts, P., & O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360–8367.

Hebscher, M., & Gilboa, A. (2016). A boost of confidence: The role of the ventromedial prefrontal cortex in memory, decision-making, and schemas. *Neuropsychologia*, *90*, 46–58.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, *35*(2), 209–223.

Jiang, F., & Luo, D. (2024). Implementing self models through joint-embedding predictive architecture. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

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

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Kihlstrom, J. F., Albright, J. S., Klein, S. B., Cantor, N., Chew, B. R., & Niedenthal, P. M. (1988). Information processing and the study of the self. In *Advances in experimental social psychology* (Vol. 21, pp. 145–178). Elsevier.

Kim, K., & Johnson, M. K. (2015). Activity in ventromedial prefrontal cortex during self-related processing: positive subjective value or personal significance? *Social cognitive and affective neuroscience*, *10*(4), 494–500.

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, *62*(1), 1–62.

LeDoux, J. E. (2020). Thoughtful feelings. *Current Biology*, *30*(11), R619–R623.

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*(10), E2016–E2025.

LeDoux, J. E., & Lau, H. (2020). Seeing consciousness through the lens of memory. *Current biology*, *30*(18), R1018–R1022.

Mobbs, D., Adolphs, R., Fanselow, M. S., Barrett, L. F., LeDoux, J. E., Ressler, K., & Tye, K. M. (2019). Viewpoints: Approaches to defining and investigating fear. *Nature neuroscience*, *22*(8), 1205–1216.

Prebble, S. C., Addis, D. R., & Tippett, L. J. (2013). Autobiographical memory and sense of self. *Psychological bulletin*, *139*(4), 815.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565–573.

Vaidya, A. R., & Badre, D. (2020). Neural systems for memory-based value judgment and decision-making. *Journal of cognitive neuroscience*, *32*(10), 1896–1923.