Trends in
Cognitive Sciences

CellPress
OPEN ACCESS

Spotlight

Emotion and prediction errors: which ingredients matter?

Marius C. Vollberg[1,2],
Yoann Stussi[2,3],
Eva R. Pool[2,3], and
David Sander[2,3,*]

Does including emotions improve reinforcement-learning models? A recent EEG study by Heffner and colleagues presents separate neural signatures for reward and emotion prediction errors. This advance invites questions about, and even holds clues to, which ingredients of emotion and prediction errors most improve reinforcement-learning models.

Reinforcement-learning models feature prominently in cognitive science and have advanced our understanding of adaptive behavior in animals, machines, and humans. These models typically center on deviations from expectations concerning reward outcomes (i.e., reward prediction errors). However, what if humans additionally use emotions about outcomes (expected and/or perceived) to learn and adapt their behavior?

Heffner and colleagues [1] recently explored the added value of including emotions in reinforcement-learning models of social behavior. Their study builds on behavioral evidence suggesting that agents' emotion prediction errors (i.e., how an outcome feels versus how it was expected to feel) may explain their behavior above and beyond what can be explained by reward signals alone [2,3]. Notably, while all reward is ultimately subjective and arguably affective in nature, here 'reward' refers to objective outcomes in the environment. Similarly, while all emotion comprises several interacting components (see below), here 'emotion' refers to valence and arousal ratings reported by participants.

Leveraging EEG, Heffner and colleagues [1] reveal signals associated with emotion prediction errors, along with changes in how those signals link to social choice behavior over time. Recent research using fMRI was unable to detect corresponding patterns when probing neural signals of each prediction error type at different timepoints: emotion prediction errors appeared different from reward prediction errors in brain-wide multivariate analyses but showed no decodable signatures themselves [4]. Probing neural signatures of prediction error types at the same timepoint, Heffner and colleagues move beyond this limited separation at the neural level by identifying distinct correlates of reward prediction errors and emotion prediction errors. In doing so, they present a step toward establishing unique contributions of emotional learning signals.

These findings suggest that self-reported emotions capture something that conventional reinforcement-learning approaches do not, thereby demonstrating the value of integrating emotion into models of cognition and behavior [5]. Given that emotions and prediction errors comprise multiple constituent parts, the findings raise questions, and offer hints, as to which ingredients are critical for improving reinforcement-learning models.

Despite substantial divergence, emotion theories [6] converge on emotion comprising more than valence. Consistent with this view, and rooted in the 'core affect' tradition, Heffner and colleagues additionally measured arousal. Unlike valence prediction errors, arousal prediction errors did not appear to explain unique variance (i.e., beyond reward and valence prediction errors) in behavior or EEG signals. This resonates with both mixed evidence in previous work and the growing consensus that the construct of arousal might be of relatively limited use [7]. From a 'core affect' perspective, absence of unique arousal associations may still be important if arousal is considered essential to 'emotion'. However, since all emotion theories feature valence [6], this result can also encourage moving beyond such definitional considerations to incorporate perspectives offered by other theories.

Emotion theories offer more than definitions; they can also help anticipate the processes that measurement might tap into [6]. Heffner and colleagues found that valence prediction errors correlate with a specific event-related potential (P3b), which in turn correlates with behavior. This intriguing connection between social behavior and neural signatures of emotion prediction errors opens new avenues for research regarding the constituent parts of emotion. Specifically, efforts to include emotions in reinforcement-learning models may benefit from considering that emotions can be viewed as comprising several distinct, interacting components [6]. Each of the five major emotion theories assumes multiple components but primarily focuses on one: while 'core affect' theories, as representatives of constructionist theories of emotion, focus on the feeling component, appraisal theories focus on cognitive appraisals, basic emotion theories on expression, motivational theories on action tendencies, and bodily/interoceptive theories on autonomic activity.

Each of these components may manifest in self-reported emotions and, therefore, underlie the EEG effects observed by Heffner and colleagues. For example, appraised relevance has been shown to shape learning from prediction errors independent of valence [8]. What appears to drive adaptation is that a given target is concern relevant, not (only) whether its valence is appraised as negative or positive.

Notably, this aligns with Heffner and colleagues' results showing that unsigned, but not signed, valence prediction errors predict P3b amplitude in a jointly estimated model. This suggests absolute differences in affect drive effects, not whether those differences were negative or positive. Thus, these results lay important groundwork for identifying the components of emotion most central to learning and behavior.

Another important implication of Heffner and colleagues' work concerns the constituent parts of prediction errors: predictions and outcomes. Self-reported emotions differ from reward in that they introduce a particular quality [9]: emotion predictions and emotion outcomes (i.e., emotion experiences) are both reported by the participant, such that outcomes may already incorporate predictions along with deviations from them. It is still intuitive to compare emotion and reward at the level of prediction errors, because of how central prediction errors have been to research on reward. However, even on the reward side, seemingly well-established prediction error effects may appear in a different light once relevant parts are adequately isolated [10].

Heffner and colleagues show that reward prediction errors and reward outcomes, when included in the same model, jointly predict behavior. The same approach yields different results for emotions: emotion (i.e., valence) outcomes alone predicted behavior (see Supplementary Table S2). As the authors note, this might be because emotion outcomes may (partially) incorporate predictions and, thus, prediction error signals. If emotion outcomes do not incorporate prediction errors, this result suggests that the variation in behavior attributed to emotion prediction errors does not primarily reflect prediction error signals per se, but rather their correlation specifically with the emotion outcome itself.

A key motivation for comparing prediction error types is the putative link between emotion prediction errors and behavior, which may serve to improve reinforcement-learning models. If emotion (or only valence) outcomes alone are most critical, or even sufficient for improving those models, that would also be a valuable insight. Thus, insights gained from joint consideration of prediction errors and their constituent parts at the behavioral level could motivate the same consideration at the neural level.

Separate neural correlates of reward and emotion prediction errors advance a burgeoning literature aimed at improving reinforcement-learning models by including emotion. Future advances building on this work will be well positioned to rigorously test constituent parts of emotion and prediction errors from the outset.

Declaration of interests

The authors declare that they have no conflicts of interest.

[1]Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Zurich, Switzerland
[2]Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
[3]Department of Psychology, FPSE, University of Geneva, Geneva, Switzerland

*Correspondence:
David.Sander@unige.ch (D. Sander).

https://doi.org/10.1016/j.tics.2025.10.006

References

1. Heffner, J. *et al.* (2025) Separable neural signals for reward and emotion prediction errors. *Nat. Commun.* 16, 7849
2. Heffner, J. *et al.* (2021) Emotion prediction errors guide socially adaptive behaviour. *Nat. Hum. Behav.* 5, 1391–1401
3. Vollberg, M.C. and Cikara, M. (2024) Affective prediction errors in persistence and escalation of aggression. *J. Exp. Psychol. Gen.* 153, 1551–1567
4. Xu, T. *et al.* (2025) Distinct neural computations scale the violation of expected reward and emotion in social transgressions. *Commun. Biol.* 8, 106
5. Dukes, D. *et al.* (2021) The rise of affectivism. *Nat. Hum. Behav.* 5, 816–820
6. Sander, D. (2025) Theories of emotion for human affective neuroscience. In *The Cambridge Handbook of Human Affective Neuroscience* (2nd edn) (Armony, J.L. and Vuilleumier, P., eds), pp. 7–32, Cambridge University Press
7. Smith, K.E. *et al.* (2025) Arousal may not be anything to get excited about. *Emot. Rev.* 17, 3–15
8. Stussi, Y. (2025) Affective and computational determinants of threat extinction biases. *Behav. Res. Ther.* 192, 104804
9. Vollberg, M.C. and Sander, D. (2024) Hidden reward: affect and its prediction errors as windows into subjective value. *Curr. Dir. Psychol. Sci.* 33, 93–99
10. Feher Da Silva, C. *et al.* (2023) Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nat. Hum. Behav.* 7, 956–969