Published in final edited form as:
Nat Hum Behav. 2021 October ; 5(10): 1391–1401. doi:10.1038/s41562-021-01213-6.

Emotion Prediction Errors Guide Socially Adaptive Behavior

Joseph Heffner[1], **Jae-Young Son**[1], **Oriel FeldmanHall**[1,2]

[1]Department of Cognitive, Linguistic, Psychological Sciences, Brown University, Providence RI 02912

[2]Carney Institute for Brain Science, Brown University, Providence RI 02912

Abstract

People make decisions based on deviations from expected outcomes, known as prediction errors. Past work has focused on reward prediction errors, largely ignoring violations of expected emotional experiences—emotion prediction errors. We leverage a method to measure real-time fluctuations in emotion as people decide to punish or forgive others. Across four studies (N=1,016), we reveal that emotion and reward prediction errors have distinguishable contributions to choice, such that emotion prediction errors exert the strongest impact during decision-making. We additionally find that a choice to punish or forgive can be decoded in less than a second from an evolving emotional response, suggesting emotions swiftly influence choice. Finally, individuals reporting significant levels of depression exhibit selective impairments in using emotion—but not reward—prediction errors. Evidence for emotion prediction errors potently guiding social behaviors challenge standard decision-making models that have focused solely on reward.

Introduction

How do we learn to make adaptive decisions, such as whether to avoid a risky financial endeavor or start a collaboration with a new colleague? A rich literature on value-based decision-making illustrates that choices are made based on the expectation of rewards, and that violations of these reward expectations—i.e., prediction errors—enable an agent to update their knowledge about their environment to facilitate survival[1–4]. Over the last few decades, these insights have been elegantly encapsulated in a reinforcement learning framework[5], which has served as the foundation for virtually all standard models of decision-making. Even complex social behaviors, such as affiliating with coworkers or reconciling with a spouse, are thought to be motivated by the violation of expected outcomes[6,7]. To illustrate, a colleague's failure to meet a deadline might generate a negative prediction error, which in turn drives learning through continued reinforcement (e.g., this

Corresponding Author: Oriel FeldmanHall, Brown University, 190 Thayer St., Providence, RI 02912, oriel_feldmanhall@brown.edu.

Author Contributions Statement
J.H., J.S., and O.F.H. contributed to designing the research and writing the paper. J.H. analyzed the data.

Competing Interests Statement
The authors declare no competing interests.

colleague is often late to meetings) and adjustments of future behaviors (e.g., collaborations with this colleague are to be avoided).

In parallel with research linking reward to decision-making, a separate literature also demonstrates that emotion exerts a powerful influence on choice[8–12]. Although there has been interest in understanding how anticipated emotions affect behavior[13–18], relatively little work has examined how the violation of expected emotions—a concept we label emotion prediction errors—influence decision-making, especially in the context of social interactions[19–21]. A person may, for example, avoid collaborating because she dreads interactions with her aloof colleague, only to find out that once the collaboration begins, her colleague is actually quite warm and humorous. The unexpected joy of working with this colleague therefore produces a positive emotion prediction error, which motivates more extensive future collaborations. Prior work shows that sophisticated mental models of emotion are used to predict how other people transition between distinct emotional states[22], and predictions about expected aversive emotions such as regret and guilt can shape social interactions[23–25]. However, whether violations of expected emotions also affect decision-making is an open question[19,26].

Additionally, little is known about how reward and emotion relate to one another, as past work on decision-making has either ignored emotional experiences, or else assumed that emotion is synonymous with reward value[2,27–31]. For example, in a reinforcement learning framework, external rewards (e.g., money, juice) are used to update an agent's value function, and any state changes (such as emotions) are considered to be nuisance variables[32]. Other accounts hint that emotions are simply an internal proxy for value, such that emotions may shape how an individual processes the subjective value of a choice by applying a (nonlinear) transformation to objective reward[33]. This lack of consensus and clarity impacts the specificity of theories of decision-making and hampers insight into a variety of psychopathologies that are canonically associated with deficits in both reward and emotion processing[34–37]. For instance, it has not been determined whether emotion and reward independently or jointly impact socially maladaptive behaviors accompanying mood disorders, such as depression[38]. Therefore, in order to gain a holistic understanding of the mechanisms guiding adaptive social decision-making[39,40], it is critical to map the relative contributions of reward and emotion prediction errors to behavior.

To test how strongly reward and emotion prediction errors impact social behaviors like punishing or forgiving others, we quantify how violations of emotional expectations bias choices in multiple interactive economic games. As a direct analogue to reward prediction errors, we examine emotion prediction errors using a framework that treats emotion by its basic psychological constituents free of any implied cognitive structures[41]. This model of emotion partitions emotional experiences into the affective dimensions of valence (pleasurableness) and arousal (alertness/activation)[42,43], which jointly constitute the core affect of emotion (we have adopted the term "emotion prediction errors", rather than affective prediction errors, as it captures the conscious emotional experiences participants are being asked to measure and report during these tasks.). We developed a technique that measures real-time fluctuations in emotions as the decision-process unfolds, enabling us to precisely and mathematically map the subjective experience of emotion alongside economic rewards during social exchanges. This allows us to test the possibility that violations of emotion expectations influence socially-consequential choices, such as deciding to punish or forgive a norm violator. In the tradition of reinforcement learning, we consider reward as an external reinforcer, such as money or food, that encourages similar future behaviors[32,44–49]. This allows us to compare the relative strengths of reward and emotion prediction errors on choice, while remaining agnostic about whether (and/or how) these prediction errors may eventually be integrated into a common value signal reflecting 'net value'[50].

Across four separate experiments, participants (N=1,016) played one of two behavioral economic games—the Ultimatum Game[51] or Justice Game[52]—while simultaneously rating their affective experiences using a measure we term *dynamic Affective Representation Mapping* (dARM), adapted from the affect grid used in past research[43]. This measure represents a subjective map of emotional responses where the horizontal axis characterizes the valence dimension, and the vertical axis characterizes the arousal dimension. A person who is feeling angry might, for example, report high arousal and negative valence by rating their emotional state in the upper-left corner of the grid (Fig. 1A).

In Experiment 1, participants (N=364) completed multiple rounds of a one-shot Ultimatum Game (UG) online, which captures punitive responses to fairness violations in a dyadic social interaction. Using a between-subjects design, participants played either as the Responder or a third-party making decisions on behalf of an anonymous Responder. In the UG, the Responder received an unfair monetary split from the Proposer, and participants were then tasked with deciding whether to accept the Proposer's offer as-is, or else reject the offer (i.e., costly punishment such that neither the Proposer nor Responder receives any money). In our modified version, participants made ratings on the dARM at two critical time points: first, at the beginning of the trial before there was any monetary offer from the Proposer, which captures participants' emotion expectations, and second, after the Proposer makes an offer, which captures emotion experience (Fig. 1B). By using the dARM to measure emotions as a social interaction unfolds, we can mathematically compute the difference between emotion expectations and compare them to the actual emotional experience, effectively capturing emotion prediction errors (see Methods). These emotion prediction errors were measured on two dimensions, valence and arousal, such that a valence prediction error would be calculated by the difference between the predicted versus experienced (un)pleasantness of the offer, while an arousal prediction error would be the difference between predicted and experienced arousal (Fig. 1C).

Mirroring how reward prediction errors are typically treated in the literature[20,53], the effects of reward prediction errors were captured by having participants make trial-by-trial predictions about the reward they expected to receive from the Proposer, which could then be compared to the actual offer received (by subtracting the received offer from the predicted offer). This design critically allowed us to distinguish between the contributions of reward and emotion prediction errors (PEs) during a dynamic social interaction using the following conceptual model:

$$Choice \sim Reward\ PE + Valence\ PE + Arousal\ PE \tag{1}$$

Before examining how well prediction errors for reward, valence, and arousal govern decisions to punish, we scaled all PEs at the group level prior to being modeled (similar to z-scoring without mean-centering, as 0 is the meaningful case where predictions perfectly match experience), which permitted a direct comparison of their relative contributions to choice using a common metric[54,55].

Results

Emotion and reward prediction errors have distinguishable contributions to choice

Results reveal that all three types of PEs contribute to decisions to punish. Participants punished at higher rates when experiencing less reward or valence than expected, or more arousal than expected (Table 1, Fig. 2A; the same pattern of results was found using nonparametric regression, see Supplement). A likelihood ratio test demonstrated that the sequential addition of valence ($\chi^2(4) = 512.65, p < .001$) and arousal PEs ($\chi^2(4) = 70.24$, $p < .001$) significantly improved the explanatory power of the model over a more traditional analysis which only included reward PEs. Given past work that has characterized emotional valence as a byproduct of reward processing[56–58], it is particularly noteworthy that we find a unique contribution of valence prediction errors—e.g., surprisingly negative feelings such as disappointment or sadness—for punitive decisions. While the valence and reward PEs are correlated at the intra-individual level ($r_{rm} = 0.80$, $p < .001$), the Variance Inflation Factor (VIF) statistics indicate low collinearity between these predictors in our model ($VIF_{valence} = 1.55$, $VIF_{reward} = 1.54$, $VIF_{arousal} = 1.04$), and therefore produces reliable estimates of how strongly these PEs affect choice. Moreover, using a beta coefficient test[59], we found that valence PEs have a significantly stronger impact on motivating punitive choices than reward PEs ($z = -3.74$, $p < .001$). That is, while people do rely on reward PEs to inform their choices, they rely even more on negative deviations from expected emotional valence.

An alternative explanation of these findings is that emotion prediction errors are merely soaking up the additional variance that would be typically captured by modeling individual differences in the subjective valuation of reward. To test for this possibility, we pitted our empirical prediction error model against standard utility models that leverage an exponential scaling parameter to capture any non-linear valuations of rewards[33,53,60]. We used the following equations to transform objective reward magnitudes into subjective value before calculating the resulting subjective reward prediction error (sRPE):

$$\begin{gathered} sRPE = reward_{actual}^{\lambda} - reward_{prediction}^{\lambda} \\ where\ 0 \leq \lambda \leq 1 \end{gathered} \tag{2}$$

By bounding lambda, this utility model captures the diminishing marginal utility of reward in the tradition of classic utility models [33]. We incorporated subjective reward prediction errors into a utility model by adding an additional free parameter w_1, which specifies the subject's weight on model-derived sRPEs:

$$\begin{gathered} utility_{accept} = w_1 sRPE \\ utility_{reject} = 0 \end{gathered} \tag{2.1}$$

The choice rule was computed by placing the utility values for each decision into a softmax function:

$$p(accept) = \frac{e^{\beta(utility_{accept})}}{e^{\beta(utility_{accept})} + e^{\beta(utility_{reject})}} \qquad (2.2)$$

$$p(reject) = 1 - p(accept)$$

Thus, we can compare how our empirically-derived prediction errors (reward, valence, arousal) fare against subjective reward prediction errors that incorporate non-linear valuation of reward (see Supplementary Figure 4 and Supplementary Table 8). Results reveal that the prediction error model that includes empirical reward and valence (but not arousal) PEs outperforms all other models, including those that rely on model-derived reward prediction errors, t(363) =−16.09, p < .001, or other impoverished models that do not account for valence PEs. Furthermore, these results suggest that valence prediction errors in particular are not merely reflecting individual differences in the subjective valuation of reward.

While the additive model (equation 1) provides the most direct comparison of the strength of each prediction error on decisions to punish, we conducted a secondary analysis to assess whether prediction errors also exert any joint influence on choice. By testing all possible interactions between all three prediction error types in a mixed-effects regression, we found a significant three-way interaction between reward, valence, and arousal PEs (β = −0.37, SE = 0.07, $z = -4.91$, $p < .001$), a significant interaction between valence and arousal PEs (β = −0.36, SE = 0.11, $z = -3.22$, $p = .001$) and reward and arousal PEs (β = −0.30, SE = 0.11, $z = -2.65$, $p = .008$), but not between reward and valence PEs ($p = .60$; see Supplementary Table 4). Together, this suggests that the strength of a given prediction error is partially modulated by other prediction errors, such that arousal PEs appear to augment the role of valence and reward PEs.

We conducted follow-up analyses to assess the robustness of our results and check for potential nonlinearities in the data. First, we tested for nonlinear effects of prediction errors on decision-making using a generalized additive mixed effects model (GAMM) which showed that the marginal contribution of valence PEs had a stronger unique contribution to choice than reward PEs (see Supplementary Figure 3 and Supplementary Table 2). Second, we tested the strength of prediction errors by controlling for expectations (i.e., modelling both the prediction and prediction error in the same regression)[53], which revealed that valence and reward prediction errors still explain significant variance in decisions to punish, even when controlling for expectations (see Supplementary Table 5). Third, we can directly examine the contributions of rewarding and emotional experiences on decisions to punish by fitting a model that only includes information about participants' actual experiences of reward and emotion, independent of their expectations. Results reveal that experienced reward (i.e., the offer itself) predicts decisions to punish more strongly than experiences of emotional valence (beta-comparison $z = -4.37$, $p < .001$) or arousal (beta-comparison $z = -10.57$, $p < .001$). Finally, since querying emotion predictions directly after reward predictions could have diminished the role of reward PEs, we ran a subsequent pre-registered experiment to replicate our findings while controlling for potential ordering effects (https://osf.io/3mgxz/). In Experiment 2 (N = 228), there was no evidence that reward PEs were dampened by the presence of asking participants to predict and report their emotional experiences (see Supplementary Figure 7 and Supplementary Tables 9 – 11). Moreover, as observed in Experiment 1, valence PEs had the strongest impact on shaping decisions to punish compared to reward PEs (beta comparison: $z = -2.57$, $p = .005$).

These results have three important implications. First, reward and emotion expectation violations are distinguishable inputs during decisions to punish. Second, although reward prediction errors have traditionally been treated as the predominant driver of punitive decisions in social exchanges[61,62], our findings instead indicate that valence emotion prediction errors are actually the strongest motivator. Third, when considering the direct contributions of experienced reward and emotion, reward appears to more strongly bias behavior than emotion, illustrating that emotion only outperforms reward once prediction errors are considered.

These findings demonstrate a link between emotion prediction errors and punishment, suggesting that violations of emotion expectations are integral to motivating social choice. It remains unclear, however, how emotions are constructed during the decision-making process, or when these emotional experiences ultimately bias choice. Even when the contributions of emotion are considered alongside reward, emotions are typically treated as a static input rather than a dynamic process[20,63]. This assumption places artificial constraints on the role emotions play in biasing choice, and is incongruent with theoretical accounts claiming that dynamic fluctuation is a core feature of emotion[16]. Indeed, most major theories of emotion propose that changes in the intensity of experienced affect over time can be integral in shaping behaviors[64,65], and can adaptively vary depending on cue relevant environmental changes[66].

Temporal dynamics of emotion and its relationship with choice

Therefore, to gain a more granular perspective of how emotion biases choice, and to test the robustness of the emotion prediction error effect observed in Experiments 1 and 2, we conducted a third experiment exploring the temporal dynamics of emotion while simultaneously employing a stronger experimental control for the influence of reward on social choice. One of the limitations of using the UG to study dynamic changes in reward and emotion is that each of the options (i.e., Accept or Reject) results in different monetary reward. However, if the relative contribution of reward were experimentally held constant, this would allow us to decouple monetary reward from emotion, and more directly examine how emotion prediction errors influence choice. Therefore, to control for the variable influence of reward, in Experiment 3 participants (N=73) played a modified version of the UG called the Justice Game (JG) in a laboratory setting[52]. In the JG, participants always played as the Responder and received unfair offers from various Proposers. After receiving an offer, Responders could redistribute the money between themselves and the Proposer. Analogously to the UG, two of the redistribution options were to Accept (take the offer as-is) or Punish (reduce Proposer’s payoff to match the Responder’s). The JG introduces two additional unique options which capture preferences between non-punitive responses and cost-free punishment: Responders could non-punitively Compensate by increasing their own payout to match the amount of money Proposers kept for themselves or apply a cost-free punishment by Reversing the proposed payoffs, such that Proposers get what they offered to Responders (see Methods). On each trial, only two of the available four options were randomly presented, and participants were never aware which two options would be available. This structure was important for two reasons. First, because there was uncertainty regarding which choice pair participants would receive, the final monetary outcomes were experimentally decoupled from the Proposer's original offer. This provided an ideal testbed for studying continuous fluctuations in emotions, as participants' emotions could change over the course of a trial as they received new information about Proposers' offers and the possible ways to redistribute the money. Second, by matching the participant's payout by holding reward constant when certain choices were presented together (i.e., Compensate/ Reverse and Accept/Punish), the task structure provides a strong experimental test of how emotion influences social choice independently from the final reward outcome.

As before, we used the dARM to measure participants' emotion predictions about the Proposer's offer and emotional experiences after receiving the Proposer's offer. Because the available choice pair was unpredictable, we additionally measured participants' emotional experiences after making their decision. All emotion measurements were sampled every 10ms using mouse tracking, which allowed us to continually measure emotion predictions and experiences as they unfolded in real-time.

We first calculated emotion prediction errors using participants' final emotion rating (mirroring the analysis performed in Experiments 1–2). Results reveal that negative valence emotion prediction errors robustly predict choices to Punish ($\beta = -1.26$, SE = .24, z = −5.18, $p < .001$) and Reverse ($\beta = -0.97$, SE = .26, z = −3.74, $p < .001$). Although significantly predictive of punitive behavior, these valence emotion prediction errors did not outperform reward prediction errors for either choice pair (Accept/Punish: $z = -.52$, $p = .30$, Fig. 2B; Compensate/Reverse: $z = -.43$, $p = .34$, Fig. 2C). In contrast to Experiments 1–2, arousal prediction errors were not significantly predictive of decisions to Punish or Reverse (all Ps>.83, see Supplementary Tables 12 and 13), suggesting that the strength of the arousal PE signal may be context dependent. In addition, we tested how these empirically-derived prediction errors fare against subjective (model derived, detailed above in [equation 2]) reward prediction errors. We found that a model that includes empirical reward and valence —but not arousal—PEs outperforms all other models, including those that only rely on model-derived reward prediction errors (*$t(72) = -49.7$, $p < .001$*; see Supplementary Figure 8 and Supplementary Table 15). Taken alongside the results from Experiments 1–2, these findings suggest that emotion prediction errors—and in particular valence PEs—play a unique role in biasing various types of social choices, are at least equally potent as reward prediction errors in motivating social behaviors, and can be more powerful than reward prediction errors in certain contexts.

To further examine how the construction of emotion biases choice, we probed real-time fluctuations in participants' emotional experiences. Participants were permitted to report their emotional experiences at their own pace, resulting in trials of different lengths. We therefore resampled all emotion trajectories to a normalized timescale consisting of 100 timepoints to aid interpretation[67], and then averaged participants' emotional responses across choices (Fig. 3). The figure shows the unique emotion trajectory across time, separately for valence and arousal, for any given decision (Accept, Punish, Reverse, Compensate). When comparing the emotional trajectories in Compensate versus Reverse trials only, results reveal that participants' eventual decisions to Reverse an unfair offer —a retributive eye-for-an-eye response—can be predicted by the 37th timepoint on the normalized scale, which corresponds to 1.65s on an absolute timescale. Early emotional trajectories were so potent that we could even predict some decisions as early as half a second on an absolute timescale (for the Accept/Compensate pair with a moderately unfair offer; see Supplementary Table 16). We also examined how individuals' choices alter their emotions after the social interaction. While on average everyone's emotional valence increased after making a choice, those who responded punitively (Reverse or Punish) rapidly reported feeling positive emotional valence (Fig. 3)—a "joy of punishment" effect. Together, these results reveal that by experimentally stripping away the potential influence of monetary reward on choice, there is a striking impact of early emotional experiences on guiding which subsequent choice is taken—which provides further evidence that reward and emotion have unique inputs to social choice.

Functional dissociation between reward and emotion prediction errors

The privileged role of emotion prediction errors in guiding social behavior has important implications for mood disorders such as depression, which is often characterized by impairments in both reward and emotion processing[36]. To date, however, extant research on depression has examined reward and emotion in a siloed manner[68,69]—that is, they have not been interrogated side-by-side within the same paradigm—and there have been few attempts to link them to decision-making in lockstep. Consequently, it remains unclear whether reward or emotion deficits are the primary contributor to symptoms and socially maladaptive behaviors seen in clinical populations, such as anhedonia[35] and avolition[70]. We therefore conducted a preregistered fourth experiment (https://osf.io/qfejk/) comparing healthy controls against individuals reporting significant levels of depressive symptoms.

Experiment 4 (N=351) measured participants' predictions and experiences about reward and emotion in the Ultimatum Game. After completing the task, participants completed questionnaires indexing various mood disorders, including the Center for Epidemiologic Studies Depression Scale (CES-D)[71]. Participants were classified as at risk for depression or a healthy control based on scoring guidelines of depression symptomatology of the CES-D (see Methods for details about the CES-D and all additional measures).

We first observed that compared to healthy controls, those at risk for depression were less sensitive to the offer unfairness, which led them to be more punitive for fair offers and less punitive for unfair offers (Table 2). Given these observed behavioral differences, our primary goal was to then examine whether participants at risk for depression demonstrated aberrant use of reward and/or emotion prediction errors when deciding to punish. Replicating our findings from the first three experiments, healthy controls (N=205) relied most heavily on valence prediction errors when making punitive decisions ($\beta = -2.08$, SE = 0.29, z = −7.15, $p < .001$). Valence PEs were also more predictive of decisions to punish than reward PEs ($\beta = -1.41$, SE = 0.23, z = −6.03, $p = < .001$; beta coefficient test $z = -1.80$, $p = .036$;

Fig. 4A). In contrast, individuals at risk for depression (N=146) demonstrated no reliance on arousal PEs (β = 0.01, SE = 0.13, z = 0.09, *p* = .93), and significantly more reliance on reward compared to valence PEs (Reward β = −1.43, SE = 0.18, z = −7.79, *p* < .001; Valence β = −0.94, SE = 0.16, z = −5.81, *p* < .001; beta coefficient test *z* = −1.98, *p* = 0.02; Fig. 4A)—which accords with previous work showing that people with depression exhibit intact reward prediction errors in certain contexts[72].

Using the healthy controls as a benchmark, those at risk for depression exhibited selective impairment in their use of both emotion prediction errors when punishing, but there were no observable differences in reward prediction error processing. Remarkably, the attenuated reliance on valence and arousal prediction errors led to less punishment of a transgressor compared to healthy controls (Table 3; Fig. 4B). While it is possible that participants at risk for depression could simply have less reliable emotion prediction errors, this explanation is unlikely given the nearly identical distribution of arousal and valence prediction errors between groups (see Supplementary Figure 13). These findings reveal a functional dissociation between emotion and reward during the decision-making process in depression, suggesting that the two may be cognitively separable.

To further probe why participants at risk of depression relied less on emotion prediction errors, we next examined participants' responses in an independent emotion classification task. In this task, participants rated 20 canonical emotion labels (e.g., anger) on the dARM, which required them to draw upon their past memories and knowledge of how they experience each of these emotions (see Methods). Results reveal that individuals classified as depressed have a smaller range of emotional experiences (Welch two sample *t*-test, *t*(263.34) = 4.33, *p* < .001, Hedge's *g* = 0.49; Fig. 4C, 4D). Restricted emotion representations were observed along both the valence and arousal dimensions. This suggests that depression may be linked with impairments in how emotional experiences are represented[73,74], and may help explain why depression attenuated the influence of emotion prediction errors on decisions to punish in our experiment.

Discussion

Historically, there have been two major perspectives concerning the relationship between emotion, reward, and decision-making. Either emotion has been considered irrelevant or purely incidental to choice, or else emotion and reward have been treated as so intrinsically intertwined that they cannot be disentangled, thus serving similar functional roles[27–29,31]. Here, we examine the relationship between emotion and reward, revealing that neither of these accounts is accurate. By classifying emotions into distinct affective components (valence and arousal), we interrogated the possibility that emotion and reward prediction errors make both unique and interactive contributions to socially-consequential choices, such as punishing or forgiving moral transgressors. Our findings document that people rely heavily on violations of their emotional expectations to make social decisions, and that these emotion prediction errors are just as powerful, if not more potent, than reward prediction errors in guiding social behaviors.

By mathematically computing the consciously accessible affective states using the dARM, we were able to measure prediction errors for different dimensions of emotion[43], explore whether specific types of expectation violations are especially consequential for social decisions, and broaden the scope of research on anticipated emotion by creating a generalizable framework that does not rely on discrete emotions such as guilt or regret[75]. These findings build on a rich literature documenting how decision-making is influenced by emotion[8,9,11,12,28]. We demonstrate that negative valence prediction errors (negative surprise) and positive arousal prediction errors (experiencing more arousal than expected) increase the likelihood of punishing, such that valence and arousal PEs have independent and opposite effects on choice. We note, however, that the influence of arousal PEs on choice appears context-dependent, as we did not observe an effect of arousal PEs in Experiment 3. In contrast to past accounts[20,28], these results imply emotions ought to be considered in relation to the violation of an emotional expectation—not just in the emotional experience itself.

Even when the emotional experience is considered in isolation, methodological advances measuring moment-to-moment emotional changes can document the real-time evolution of how this process unfolds, clarifying emotion’s role in social decision-making. For example, we find that early emotional reactions—those that come online in less than a second during the social interaction—quickly and powerfully predict what social choices people subsequently make. Furthermore, the choices people make can drastically influence their emotional states: Choosing to punish perpetrators results in a rapid positive boost (e.g., ‘a joy of punishment’) in the wake of their decision. Together, these results accord with a growing literature on predictive processing[76], and suggests that early, transient emotional states during an unfair social exchange are essential in governing whether people ultimately decide to punish or forgive a perpetrator.

By adopting a prediction error framework to explore how social decisions are shaped by violations of emotional expectations, we were able to compare the strength of reward and emotion prediction errors in motivating decisions to punish and help others. Foundational work in decision-making has focused on rewards as external reinforcers, illustrating that reward prediction errors are an important mechanism for enabling adaptive behavior, as they allow people to compare their expectations against their experiences to modify actions accordingly[48,61]. We observe this in our own studies. Reward prediction errors explain significant variation in social behaviors and critically contribute to the choices people make during social interactions. In a similar vein, people rely on violations of their emotional expectations to calibrate their choices, and we observed a particularly robust role of valence PEs in predicting social choices across contexts. Arousal PEs, on the other hand, seem to be more sensitive to the social context, as they were not universally deployed across all experiments (i.e., the arousal PE effect was not observed in the JG, and once predictions were accounted for in the UG, arousal PEs no longer provided significant predictive value). Our computational model further hints that regardless of context, some individuals may not rely on arousal PEs at all to inform their choices. When taken together, our results suggest that the different types of prediction errors uniquely and interactively contribute to social choice, such that neither emotion nor reward predictions alone tell the whole story.

Rather, social choices appear to be the result of joint inputs from both emotion and reward prediction errors.

These findings are compatible with the theory that value is neurally encoded as a common currency, where the value of choice options under consideration are mapped to a single scale for comparison[50,77]. For example, the multiple different types of prediction errors measured in our studies—reward, valence, and arousal—may feed into an integrated value signal in the prefrontal cortex[78,79]. It additionally remains unknown whether a common value representation places equal weight on each kind of prediction error, or whether value is asymmetrically biased by valence prediction errors. Future work can help clarify how (and where) these distinct emotion and reward prediction errors are processed in the brain and the extent to which they are separable.

The adaptive qualities of emotion prediction errors become readily apparent when considering people at risk for depression, who were selectively impaired in using emotion —but not reward—prediction errors when making social decisions. We observed that individuals reporting significant levels of depression exhibited attenuated use of valence prediction errors and did not rely on arousal prediction errors at all, which led to less punishment of a norm transgressor. In contrast, they exhibited fully intact use of reward prediction errors, which accords with past research[80] (although this may be contingent on the learning context; cf.[81–83]). This pattern of relying on emotion prediction errors rather than reward prediction errors, seems central to healthy and adaptive social decision-making. Depression was associated with a reduced range of emotional responses in our studies, highlighting that emotional processes are fundamentally altered in mood disorders[73,84]. This emotional constraint may help explain the aberrant use of emotion prediction errors in those suffering from depression. Put simply, if a person is less sensitive in distinguishing between affective states, then they may be less able to choose appropriate actions given the social dynamics of the situation.

For close to a century, psychologists have sought to understand the essential drivers of human behavior. One successful framework, reinforcement learning, has elegantly illustrated that people make consequential decisions based on violations of expected rewards. This canon of work has laid the building blocks of how we understand human learning and decision-making. By adopting a similar approach, we reveal that violations of emotional expectations—emotion prediction errors—play an outsized role in guiding social behaviors. Using a variety of multimodal techniques, we document consistent evidence that these emotion prediction errors exert a strong influence on social behaviors, above and beyond reward prediction errors. Although past research has often placed reward prediction errors at the heart of decision-making, our results instead suggest that people robustly use violations of their emotional expectations to make decisions that influence both themselves and others. Contrary to conventional wisdom, we find that the only time reward plays a stronger role than emotion in decision-making is when experiences are considered in isolation from expectations. Together, these results highlight the critical importance that violations of expected emotions play, suggesting that emotional processes are just as consequential—if not more so—than violations of reward for guiding social behaviors.

Methods

Participants.

Across all four experiments, participants (N = 1,016) received either monetary compensation or partial course credit, and provided informed consent in a manner approved by Brown University's Institutional Review Board under protocol 1607001555. In Experiment 1, we aimed to collect a sample of 350 participants, which exceeds the sample sizes used in similar paradigms using reward prediction errors to study decision-making in the Ultimatum Game[20,85]. We recruited 398 individuals online using Amazon Mechanical Turk (AMT). To protect against data contamination from bots posing as real participants[86], we excluded 34 individuals' data using a conservative measure of noncompliance on the emotion classification task, which involved correctly rating the 'neutral' feeling that we explicitly instructed ought to be in the center of the dARM (see Supplementary Methods). This resulted in a final sample of 364 participants (172 female; mean age = 33.77, SD = 9.97). In Experiment 2, we collected 244 participants and excluded 16 individuals due to noncompliance, resulting in a final sample of 228 participants based on our preregistration (127 female; mean age = 35.30, SD = 11.8; see Supplement). In Experiment 3, we recruited 75 individuals and excluded 2 individuals due to noncompliance, resulting in a final sample of 73 participants (39 female; mean age = 20.33, SD = 3.27), comparable to similar paradigms using the Justice Game[52]. In Experiment 4, we aimed to collect a sample of 150 participants with depression using AMT, as detailed in our preregistration report, and we accordingly recruited a total of 508 participants. Using the preregistered exclusion criterion (identical to the one used in Experiment 1), we excluded 157 individuals from analysis due to noncompliance, resulting in a final sample of 351 participants (149 female; mean age = 35.13, SD = 10.21) with 205 healthy controls and 146 individuals classified as at risk of depression.

General procedure.

In all experiments, participants used the dARM measure to rate their emotion experiences in real time during an emotion classification task and a behavioral economic game. After completing these tasks, participants responded to a series of individual measures and/or clinical battery, depending on the experiment.

Dynamic Affective Representation Mapping (dARM) measure.

In Experiments 1, 2, and 4, we collected data using the Qualtrics online survey platform. Adapted from the affect grid used in past research[43], participants were presented with the dynamic Affective Representation Mapping (dARM) measure with a sampling resolution of 500 × 500 pixels, and asked to make their affective rating by clicking anywhere in the grid space. This enabled us to simultaneously capture fine-grained self-reports of both the valence and arousal dimensions. To familiarize participants with the use of the dARM, all participants first completed an emotion classification task. In this task, participants were asked to make ratings of 20 canonical emotion words on the grid (e.g., angry, sad, and surprised) in a randomized order. While this affective representation is typically inferred from pairwise similarity ratings of discrete emotions[87], simply training participants to interpret this subjective map has shown strong convergent validity with other approaches for emotion ratings[43]. In order to capture real-time mouse-tracking in Experiment 3, which was run in the laboratory, we used the Psychtoolbox library in Matlab to implement the dARM with a spatial resolution of 500 × 500 pixels and a temporal resolution of 10ms sampling. All participants first completed the emotion classification task.

Tasks.

In Experiment 1, participants played 20 rounds of the Ultimatum Game (UG) as either the Responder or a third-party. Since Responders and third-party deciders reacted to unfairness in similar ways, we collapsed across role for this analysis (see Supplementary Table 2). On each trial, participants were asked to answer the following questions: i) Predict how much reward the Responder would get (i.e., how much the Proposer would offer), within a range of $0 to $0.50; ii) Predict what emotions they expected to feel based on that reward; iii) Report their actual emotion experience upon receiving the offer; and iv) Decide whether to accept or reject the Proposer’s offer. The unfairness of the offer was drawn from a pseudo-random uniform distribution such that participants saw the full range of fair ($0.50, $0.50) to unfair ($0.95, $0.05) offers.

In Experiment 2, participants played 20 rounds of the UG as the Responder. All participants completed two blocks in a counterbalanced order, a reward-only block and a reward+emotion block. In the reward-only block, participants only made reward predictions without any emotion predictions or emotion experience ratings. The reward+emotion block design was the same as Experiment 1.

In Experiment 3, participants played the Justice Game (JG) in the laboratory. In the JG, participants always play as the Responder and are paired with a unique anonymous Proposer on every trial, who offers the Responder a split of money. On each of the 54 trials, the two options presented to the Responder are drawn randomly from the four available options, ensuring that participants do not know what their choice set will be ahead of time. The four available options are: 1) Accept: keeps the offer as-is; 2) Punish: reduces the payout of the Proposer to what was offered to the Responder; 3) Compensate: increase the Responder’s payout to match the Proposer’s; and 4) Reverse: swapping the payouts. For example, using the notation of ($Proposer, $Responder), if the offer was highly unfair ($9, $1), the four options would be Accept ($9, $1), Punish ($1, $1), Compensate ($9, $9), and Reverse ($1, $9). The unfairness of the offer was generated such that low, medium, and high unfair offers were equally likely and each offer was generated using a truncated normal distribution (the Proposer kept $5.10–6.30 for low offers, $6.90–8.10 for medium offers, and $8.70–9.90 for high offers). Participants were asked to do the following on a trial-by-trial basis: i) Predict how much reward they would receive, ii) Predict what emotions they expected to feel based on that reward, iii) Report their emotional experience upon receiving the Proposer’s offer, and iv) Make a decision about how to redistribute the money. Participants were also asked to report their emotional experience after making a decision, as to capture changes in their affective state depending on the available redistribution options. Because there were only 54 trials with six unique choice pairs and three levels of unfairness, time course analyses examining specific choice pairs (e.g., Compensate/Reverse) only include nine trials per subject.

In Experiment 4, participants played 20 rounds of the Ultimatum Game (UG) as the Responder. Otherwise, the UG design in Experiment 4 was identical to Experiment 1.

Post-Task Questionnaires.

Following these tasks, participants completed a series of individual difference questionnaires. In Experiments 1 and 4, we collected two survey measures for use as potential covariates: the Emotion Regulation Questionnaire[88] and the 20-item Toronto Alexithymia Scale[89]. In Experiment 4, participants also completed the Center for Epidemiologic Studies Depression Scale (CES-D)[71] and five survey measures to index how richly they experience reward and emotion: the Temporal Experience of Pleasure Scale (TEPS)[90], the Behavioral Inhibition and Behavioral Activation Scales (BISBAS)[91], the Snaith-Hamilton Pleasure Scale (SHAPS)[92], the Apathy Evaluation Scale (AES)[93], and the 20-item Toronto Alexithymia Scale (TAS)[89]. The primary measure of importance was the CES-D, as it allowed us to identify which participants were at risk of depression. Our hypotheses and predictions about all measures were preregistered and can be found in our OSF pre-registered report (https://osf.io/qfejk/). Analyses of the other measures are included in the Supplement.

Analysis.

Across all experiments, we used logistic mixed-effects regressions to predict participants' decisions using the lme4 package in R[94]. All prediction errors were calculated by taking the difference between the participants' experience (at the time of offer) and the participants' prediction (before the offer). To ensure that the beta coefficients from logistic regressions were comparable, we scaled (but did not mean-center) all prediction errors before entering them into the regression. We chose not to mean-center the prediction errors because zero indicates meaningful cases in which the participant's prediction matched their experience, therefore producing no error.

In Experiments 1 and 2, we used valence PEs, arousal PEs, and reward PEs to predict prosocial decisions to Accept or punitive decisions to Reject. The same regression specification was carried forward in all experiments. In Experiment 3, we emulated Experiment 1's analysis by taking the endpoints of the participants' mouse trajectories (i.e., the final valence and arousal ratings) to run separate logistic mixed-effects regressions for the Accept/Punish choice set and the Compensate/Reverse choice set. The regression for Experiment 4 additionally included terms accounting for being classified as in the depressed or healthy control group (based on CES-D scores), and the interactions between depression and all PE variants. To aid interpretation, we additionally performed separate regressions for participants who had significant levels of depressive symptoms and those who did not (i.e., a binary variable based on a CES-D threshold of 16, according to scoring guidelines).

In order to analyze real-time fluctuations in participants' emotions in Experiment 3, we discretized the time data into 10ms bins (i.e., our sampling rate). Because participants' emotional ratings were self-paced and had variable response times, we normalized each participants' response time on a trial-by-trial basis to compare across participants. Accordingly, participants' response times were rescaled from 1 (start of trial) to

101 (participant clicked response)[95,96]. Participants' valence and arousal measurements were averaged within each normalized time bin, allowing us to directly compare the distribution of valence and arousal responses for each of the choice sets (Accept/Punish and Compensate/Reverse) using one-way ANOVAs at each normalized time bin. To use a principled way of defining significant clusters of time bins, we used cluster-based permutation testing, which controls for multiple comparisons by generating null distributions of clusters that can be compared against true clusters[97–100]. This method estimates how big clusters would be if there would no differences between the groups (e.g., decisions to Compensate or Reverse). Permutation tests assume that these observation labels are exchangeable under the null hypothesis, such that if there is no difference between the groups, then the labels can be randomly shuffled without consequence. For each randomly permutated time series, the largest cluster statistic is calculated (the summation of F-statistics for the largest temporally continuous cluster), which represents the largest cluster that could appear due to chance. After repeating this process 1,000 times (which builds a null distribution of clusters), we test whether the clusters observed in our data are greater than 95% of the clusters expect by chance. This ensures we can precisely quantify how the evolution of emotions affects later decisions to be punitive or forgiving while controlling for multiple comparisons.

Data availability

Experiment materials information and all experiment de-identified data are publicly available at https://github.com/jpheffne/epe. The materials used in this study are widely available.

Code availability

Data analysis script notebooks are publicly available at https://github.com/jpheffne/epe.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Haoxue Fan for her assistance in running participants for Experiment 3. We appreciate the insightful feedback and comments from Michael Frank, Vishnu Murty, Amitai Shenhav, and Matt Nassar, on early manuscript drafts. The research was funded by a Center of Biological Research Excellence grant P20GM103645 from the National Institute of General Medical Sciences awarded to O.F.H. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Schultz W, Dayan P & Montague PR A neural substrate of prediction and reward. Science 275, 1593–1599, doi:DOI 10.1126/science.275.5306.1593 (1997). [PubMed: 9054347]
2. Schultz W & Dickinson A Neuronal coding of prediction errors. Annual Review of Neuroscience 23, 473–500, doi:DOI 10.1146/annurev.neuro.23.1.473 (2000).
3. King-Casas B et al. Getting to know you: reputation and trust in a two-person economic exchange. Science 308, 78–83, doi:10.1126/science.1108062 (2005). [PubMed: 15802598]

4. Pessiglione M, Seymour B, Flandin G, Dolan RJ & Frith CD Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442, 1042–1045, doi:10.1038/nature05051 (2006). [PubMed: 16929307]

5. Sutton RS & Barto AG Introduction to Reinforcement Learning. (MIT Press, 1998).

6. Ruff CC & Fehr E The neurobiology of rewards and values in social decision making. Nat Rev Neurosci 15, 549–562, doi:10.1038/nrn3776 (2014). [PubMed: 24986556]

7. Ho MK, MacGlashan J, Littman ML & Cushman F Social is special: A normative framework for teaching with and learning from evaluative feedback. Cognition 167, 91–106, doi:10.1016/j.cognition.2017.03.006 (2017). [PubMed: 28341268]

8. Bechara A, Damasio H, Tranel D & Damasio AR Deciding advantageously before knowing the advantageous strategy. Science 275, 1293–1295 (1997). [PubMed: 9036851]

9. Dalgleish T The emotional brain. Nature reviews. Neuroscience 5, 583–589 (2004). [PubMed: 15208700]

10. Vuilleumier P How brains beware: neural mechanisms of emotional attention. Trends Cogn Sci 9, 585–594, doi:10.1016/j.tics.2005.10.011 (2005). [PubMed: 16289871]

11. Phelps EA, Lempert KM & Sokol-Hessner P Emotion and Decision Making: Multiple Modulatory Neural Circuits. Annual Review of Neuroscience 37, 263–287, doi:10.1146/annurev-neuro-071013-014119 (2014).

12. Phelps EA Emotion and cognition: insights from studies of the human amygdala. Annu. Rev. Psychol. 57, 27–53 (2006). [PubMed: 16318588]

13. Richard R, van der Pligt J & de Vries N Anticipated Affect and Behavioral Choice. Basic and Applied Social Psychology 18, 111–129, doi:10.1207/s15324834basp1802_1 (1996).

14. Caplin A & Leahy J Psychological Expected Utility Theory and Anticipatory Feelings. The Quarterly Journal of Economics 116, 55–79, doi:10.1162/003355301556347 (2001).

15. Knutson B & Greer SM Anticipatory affect: neural correlates and consequences for choice. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 3771–3786, doi:10.1098/rstb.2008.0155 (2008).

16. Nielsen L, Knutson B & Carstensen LL Affect dynamics, affective forecasting, and aging. Emotion 8, 318–330, doi:10.1037/1528-3542.8.3.318 (2008). [PubMed: 18540748]

17. Mellers BA & McGraw AP Anticipated Emotions as Guides to Choice. Current Directions in Psychological Science 10, 210–214, doi:10.1111/1467-8721.00151 (2001).

18. Mellers B, Schwartz A & Ritov I Emotion-based choice. Journal of Experimental Psychology: General 128, 332–345, doi:10.1037/0096-3445.128.3.332 (1999).

19. FeldmanHall O & Chang LJ in Goal-Directed Decision Making (eds Richard Morris, Bornstein Aaron, & Shenhav Amitai) Ch. 14, 309–330 (Academic Press, 2018).

20. Xiang T, Lohrenz T & Montague PR Computational Substrates of Norms and Their Violations during Social Exchange. The Journal of Neuroscience 33, 1099 (2013). [PubMed: 23325247]

21. Hétu S, Luo Y, D'Ardenne K, Lohrenz T & Montague PR Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. Social Cognitive and Affective Neuroscience 12, 1972–1982, doi:10.1093/scan/nsx097 (2017). [PubMed: 28981876]

22. Thornton MA & Tamir DI Mental models accurately predict emotion transitions. Proceedings of the National Academy of Sciences 114, 5982, doi:10.1073/pnas.1616056114 (2017).

23. Chang L & Sanfey A Unforgettable ultimatums? Expectation violations promote enhanced social memory following economic bargaining. Frontiers in Behavioral Neuroscience 3, 36, doi:10.3389/neuro.08.036.2009 (2009). [PubMed: 19876405]

24. Chang LJ & Sanfey AG Great expectations: neural computations underlying the use of social norms in decision-making. Social Cognitive and Affective Neuroscience 8, 277–284, doi:10.1093/scan/nsr094 (2011). [PubMed: 22198968]

25. Loewenstein G & Lerner JS in Handbook of affective sciences. Series in affective science. 619–642 (Oxford University Press, 2003).

26. Chang LJ & Eshin J in The nature of emotion: Fundamental questions (eds Fox Andrew S, Lapate Regina C, Shackman Alexander J, & Davidson Richard J) (Oxford University Press, 2018).

27. Hartley C & Sokol-Hessner P in The Nature of Emotion (Oxford University Press, 2017).

28. Sanfey AG Social decision-making: Insights from game theory and neuroscience. Science 318, 598–602, doi:10.1126/science.1142996 (2007). [PubMed: 17962552]

29. Pessoa L How do emotion and motivation direct executive control? Trends Cogn Sci 13, 160–166, doi:10.1016/j.tics.2009.01.006 (2009). [PubMed: 19285913]

30. O'Doherty J, Kringelbach ML, Rolls ET, Hornak J & Andrews C Abstract reward and punishment representations in the human orbitofrontal cortex. Nature Neuroscience 4, 95, doi:10.1038/82959 (2001). [PubMed: 11135651]

31. Dolan RJ Emotion, cognition, and behavior. Science 298, 1191–1194 (2002). [PubMed: 12424363]

32. Juechems K & Summerfield C Where Does Value Come From? Trends in Cognitive Sciences 23, 836–850, doi:10.1016/j.tics.2019.07.012 (2019). [PubMed: 31494042]

33. Kahneman D & Tversky A Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263–291, doi:10.2307/1914185 (1979).

34. Ironside M et al. Approach-avoidance conflict in major depression: Congruent neural findings in human and non-human primates. Biological Psychiatry, doi:10.1016/j.biopsych.2019.08.022 (2019).

35. Treadway MT & Zald DH Parsing Anhedonia: Translational Models of Reward-Processing Deficits in Psychopathology. Current Directions in Psychological Science 22, 244–249, doi:10.1177/0963721412474460 (2013). [PubMed: 24748727]

36. Whitton AE, Treadway MT & Pizzagalli DA Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. Curr Opin Psychiatry 28, 7–12, doi:10.1097/YCO.0000000000000122 (2015). [PubMed: 25415499]

37. Hooker C & Park S Emotion processing and its relationship to social functioning in schizophrenia patients. Psychiatry research 112, 41–50, doi:10.1016/S0165-1781(02)00177-4 (2002). [PubMed: 12379449]

38. Pessoa L On the relationship between emotion and cognition. Nature Reviews Neuroscience 9, 148–158, doi:DOI 10.1038/nrn2317 (2008). [PubMed: 18209732]

39. Berridge KC The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191, 391–431, doi:10.1007/s00213-006-0578-x (2007). [PubMed: 17072591]

40. Berridge KC, Robinson TE & Aldridge JW Dissecting components of reward: 'liking', 'wanting', and learning. Curr Opin Pharmacol 9, 65–73, doi:10.1016/j.coph.2008.12.014 (2009). [PubMed: 19162544]

41. Russell JA Core affect and the psychological construction of emotion. Psychological Review 110, 145–172, doi:10.1037/0033-295X.110.1.145 (2003). [PubMed: 12529060]

42. Russell JA & Barrett LF Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology 76, 805–819, doi:10.1037/0022-3514.76.5.805 (1999). [PubMed: 10353204]

43. Russell JA, Weiss A & Mendelsohn GA Affect Grid: A single-item scale of pleasure and arousal. Journal of Personality and Social Psychology 57, 493–502, doi:10.1037/0022-3514.57.3.493 (1989).

44. Dayan P & Daw ND Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience 8, 429–453, doi:10.3758/CABN.8.4.429 (2008).

45. Berridge KC From prediction error to incentive salience: mesolimbic computation of reward motivation. European Journal of Neuroscience 35, 1124–1143, doi:10.1111/j.1460-9568.2012.07990.x (2012).

46. Kaelbling LP, Littman ML & Moore AW Reinforcement learning: a survey. J. Artif. Int. Res. 4, 237–285 (1996).

47. Montague PR, Dayan P & Sejnowski TJ A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of Neuroscience 16, 1936, doi:10.1523/JNEUROSCI.16-05-01936.1996 (1996). [PubMed: 8774460]

48. Schultz W, Dayan P & Montague PR A Neural Substrate of Prediction and Reward. Science 275, 1593 (1997). [PubMed: 9054347]

49. Daw ND, Niv Y & Dayan P Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8, 1704–1711, doi:http://www.nature.com/neuro/journal/v8/n12/suppinfo/nn1560_S1.html (2005). [PubMed: 16286932]

50. Levy DJ & Glimcher PW The root of all value: a neural common currency for choice. Current opinion in neurobiology 22, 1027–1038, doi:10.1016/j.conb.2012.06.001 (2012). [PubMed: 22766486]

51. Güth W, Schmittberger R & Schwarze B An experimental analysis of ultimatum bargaining. Journal of economic behavior & organization 3, 367–388 (1982).

52. FeldmanHall O, Sokol-Hessner P, Van Bavel JJ & Phelps EA Fairness violations elicit greater punishment on behalf of another than for oneself. Nat Commun 5, 5306 (2014). [PubMed: 25350814]

53. Rutledge RB, Skandali N, Dayan P & Dolan RJ A computational and neural model of momentary subjective well-being. Proceedings of the National Academy of Sciences 111, 12252, doi:10.1073/pnas.1407535111 (2014).

54. Aiken LS & West SG in Multiple regression: Testing and interpreting interactions. xi, 212–xi, 212 (Sage Publications, Inc, Thousand Oaks, CA, US, 1991).

55. Iacobucci D, Schneider MJ, Popovich DL & Bakamitsos GA Mean centering helps alleviate “micro” but not “macro” multicollinearity. Behavior Research Methods 48, 1308–1317, doi:10.3758/s13428-015-0624-x (2016). [PubMed: 26148824]

56. Delgado MR, Locke HM, Stenger VA & Fiez JA Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. Cognitive, Affective, & Behavioral Neuroscience 3, 27–38, doi:10.3758/CABN.3.1.27 (2003).

57. Murray EA The amygdala, reward and emotion. Trends in Cognitive Sciences 11, 489–497, doi:10.1016/j.tics.2007.08.013 (2007). [PubMed: 17988930]

58. Lang PJ & Bradley MM in Handbook of approach and avoidance motivation. 51–65 (Psychology Press, 2008).

59. Clogg CC, Petkova E & Haritou A Statistical Methods for Comparing Regression Coefficients Between Models. American Journal of Sociology 100, 1261–1293 (1995).

60. Sokol-Hessner P et al. Thinking like a trader selectively reduces individuals' loss aversion. Proceedings of the National Academy of Sciences 106, 5035, doi:10.1073/pnas.0806761106 (2009).

61. King-Casas B et al. Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. Science 308, 78 (2005). [PubMed: 15802598]

62. Montague PR & Lohrenz T To detect and correct: norm violations and their enforcement. Neuron 56, 14–18 (2007). [PubMed: 17920011]

63. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE & Cohen JD The neural basis of economic decision-making in the Ultimatum Game. Science 300, 1755–1758, doi:10.1126/science.1082976 (2003). [PubMed: 12805551]

64. Scherer KR Emotions are emergent processes: they require a dynamic computational architecture. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 3459–3474, doi:10.1098/rstb.2009.0141 (2009).

65. Diener E, Larsen RJ, Levine S & Emmons RA Intensity and frequency: dimensions underlying positive and negative affect. J Pers Soc Psychol 48, 1253–1265, doi:10.1037//0022-3514.48.5.1253 (1985). [PubMed: 3998989]

66. Kuppens P, Oravecz Z & Tuerlinckx F Feelings change: Accounting for individual differences in the temporal dynamics of affect. Journal of Personality and Social Psychology 99, 1042–1060, doi:10.1037/a0020962 (2010). [PubMed: 20853980]

67. Freeman JB & Ambady N MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods 42, 226–241, doi:10.3758/BRM.42.1.226 (2010). [PubMed: 20160302]

68. Admon R & Pizzagalli DA Dysfunctional reward processing in depression. Current Opinion in Psychology 4, 114–118, doi:10.1016/j.copsyc.2014.12.011 (2015). [PubMed: 26258159]

69. Keren H et al. Reward Processing in Depression: A Conceptual and Meta-Analytic Review Across fMRI and EEG Studies. Am J Psychiatry 175, 1111–1120, doi:10.1176/appi.ajp.2018.17101124 (2018). [PubMed: 29921146]

70. Dowd EC, Frank MJ, Collins A, Gold JM & Barch DM Probabilistic Reinforcement Learning in Patients With Schizophrenia: Relationships to Anhedonia and Avolition. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 1, 460–473, doi:10.1016/j.bpsc.2016.05.005 (2016). [PubMed: 27833939]

71. Radloff LS Vol. 1 385–401 (Sage Publications, US, 1977).

72. Moutoussis M et al. Neural activity and fundamental learning, motivated by monetary loss and reward, are intact in mild to moderate major depressive disorder. PLOS ONE 13, e0201451, doi:10.1371/journal.pone.0201451 (2018). [PubMed: 30071076]

73. Demiralp E et al. Feeling blue or turquoise? Emotional differentiation in major depressive disorder. Psychological science 23, 1410–1416, doi:10.1177/0956797612444903 (2012). [PubMed: 23070307]

74. Emotional complexity (The Guilford Press, New York, NY, US, 2008).

75. Chang LJ, Smith A, Dufwenberg M & Sanfey AG Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70, 560–572 (2011). [PubMed: 21555080]

76. Hutchinson JB & Barrett LF The Power of Predictions: An Emerging Paradigm for Psychological Research. Current Directions in Psychological Science 28, 280–291, doi:10.1177/0963721419831992 (2019). [PubMed: 31749520]

77. Glimcher PW in Neuroeconomics (Second Edition) (eds Glimcher Paul W. & Fehr Ernst) 373–391 (Academic Press, 2014).

78. Chib VS, Rangel A, Shimojo S & Doherty JP Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. The Journal of Neuroscience 29, 12315, doi:10.1523/JNEUROSCI.2575-09.2009 (2009). [PubMed: 19793990]

79. Smith DV et al. Distinct Value Signals in Anterior and Posterior Ventromedial Prefrontal Cortex. The Journal of Neuroscience 30, 2490, doi:10.1523/JNEUROSCI.3319-09.2010 (2010). [PubMed: 20164333]

80. Rutledge RB et al. Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. JAMA Psychiatry 74, 790–797, doi:10.1001/jamapsychiatry.2017.1713 (2017). [PubMed: 28678984]

81. Gradin VB et al. Expected value and prediction error abnormalities in depression and schizophrenia. Brain : a journal of neurology 134, 1751–1764, doi:10.1093/brain/awr059 (2011). [PubMed: 21482548]

82. Kumar P et al. Abnormal temporal difference reward-learning signals in major depression. Brain : a journal of neurology 131, 2084–2093, doi:10.1093/brain/awn136 (2008). [PubMed: 18579575]

83. Kumar P et al. Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. Neuropsychopharmacology 43, 1581–1588, doi:10.1038/s41386-018-0032-x (2018). [PubMed: 29540863]

84. Ehring T, Tuschen-Caffier B, Schnulle J, Fischer S & Gross JJ Emotion regulation and vulnerability to depression: spontaneous versus instructed use of emotion suppression and reappraisal. Emotion 10, 563–572, doi:10.1037/a0019010 (2010). [PubMed: 20677873]

85. Hétu S, Luo Y, D’Ardenne K, Lohrenz T & Montague PR Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. Social Cognitive and Affective Neuroscience, nsx097–nsx097, doi:10.1093/scan/nsx097 (2017).

86. Chmielewski M & Kucker SC An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. Social Psychological and Personality Science, 1948550619875149, doi:10.1177/1948550619875149 (2019).

87. Posner J, Russell JA & Peterson BS The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17, 715–734, doi:10.1017/S0954579405050340 (2005). [PubMed: 16262989]

88. Gross JJ & John OP Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. J Pers Soc Psychol 85, 348–362, doi:10.1037/0022-3514.85.2.348 (2003). [PubMed: 12916575]

89. Bagby RM, Parker JDA & Taylor GJ The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. Journal of Psychosomatic Research 38, 23–32, doi:10.1016/0022-3999(94)90005-1 (1994). [PubMed: 8126686]

90. Gard DE, Gard MG, Kring AM & John OP Anticipatory and consummatory components of the experience of pleasure: A scale development study. Journal of Research in Personality 40, 1086–1102, doi:10.1016/j.jrp.2005.11.001 (2006).

91. Carver CS & White TL Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. Journal of Personality and Social Psychology 67, 319–333, doi:10.1037/0022-3514.67.2.319 (1994).

92. Snaith RP et al. A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. The British journal of psychiatry : the journal of mental science 167, 99–103 (1995). [PubMed: 7551619]

93. Marin RS, Biedrzycki RC & Firinciogullari S Reliability and validity of the Apathy Evaluation Scale. Psychiatry research 38, 143–162 (1991). [PubMed: 1754629]

94. Bates D, Mächler M, Bolker B & Walker S Fitting Linear Mixed-Effects Models Using lme4. 2015 67, 48, doi:10.18637/jss.v067.i01 (2015).

95. Freeman J, Dale R & Farmer T Hand in Motion Reveals Mind in Motion. Frontiers in Psychology 2, doi:10.3389/fpsyg.2011.00059 (2011).

96. Wulff DU, Haslbeck JMB, Kieslich PJ, Henninger F & Schulte-Mecklenbeck M in A Handbook of Process Tracing Methods (eds Schulte_Mecklenbeck M, Kughberger A, & Johnson JG) (2019).

97. Maris E & Oostenveld R Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164, 177–190, doi:10.1016/j.jneumeth.2007.03.024 (2007). [PubMed: 17517438]

98. Barr D clusterperm, <https://github.com/dalejbarr/clusterperm> (2019).

99. Frömer R, Maier M & Abdel Rahman R Group-Level EEG-Processing Pipeline for Flexible Single Trial-Based Analyses Including Linear Mixed Models. Frontiers in Neuroscience 12, 48 (2018). [PubMed: 29472836]

100. Maris E Statistical testing in electrophysiological studies. Psychophysiology 49, 549–565, doi:10.1111/j.1469-8986.2011.01320.x (2012). [PubMed: 22176204]

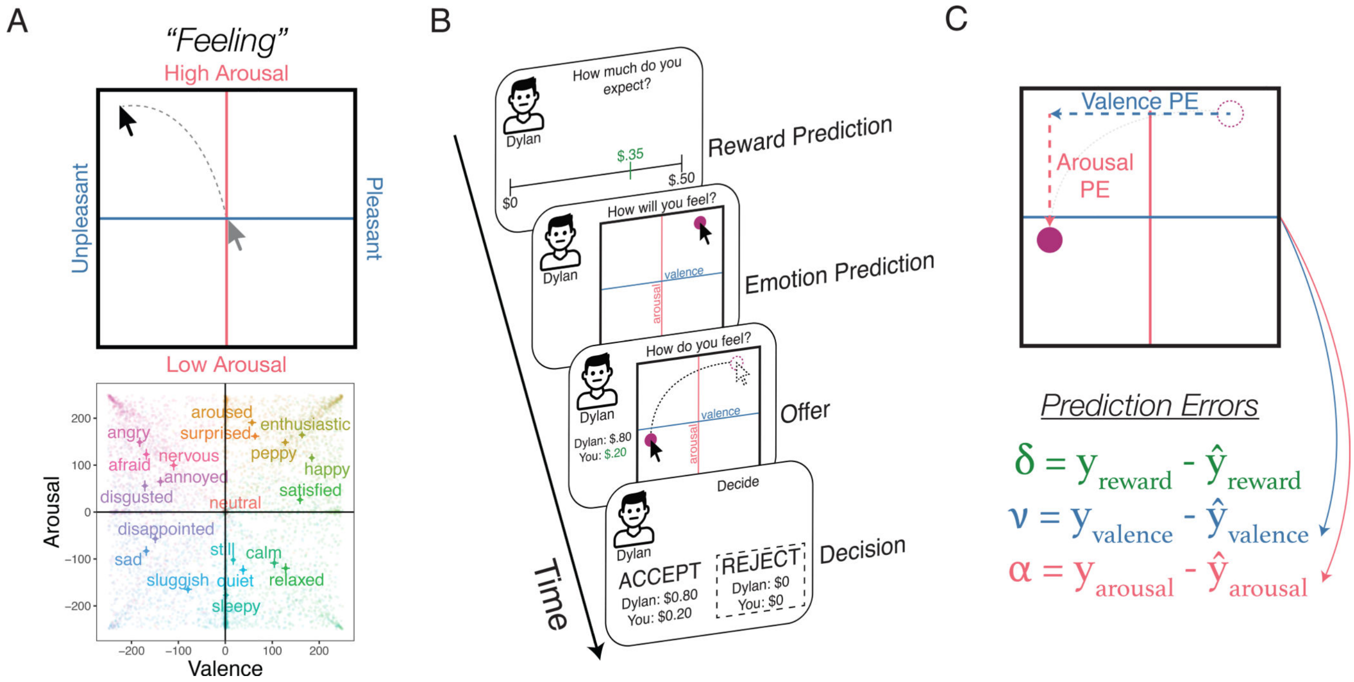

Figure 1.

A) Emotion Classification Task. Participants rate a series of 20 feelings on the dynamic Affective Representation Mapping (dARM) measure. The dARM is a 500 × 500 pixel grid, which is only delineated by a horizontal (valence axis) and vertical line (arousal axis) along with their labels. The graph below the grid shows the average ratings for 20 feeling words all participants rated in Experiment 1 (each semi-transparent datapoint reflects one individual rating). Error bars reflect 95% confidence intervals. **B) Ultimatum Game Trial Design**. Here, we show how the dARM is used in conjunction with the Ultimatum Game to capture emotion expectations and experiences. On each trial, participants make a prediction about how much money they expect to be offered, as well as a prediction about the emotions they expect to experience. Upon seeing the actual offer, participants report their current emotional experience. Finally, participants decide to either accept or reject the offer. **C) Calculating Reward and Emotion Prediction Errors.** We calculate three trial-level empirical prediction errors: a reward prediction error (δ), a valence prediction error (ν), and an arousal prediction error (α). In the equations, $\hat{y}$ refers to an individual's prediction about the reward or emotion they would experience, and y refers to their actual experience.

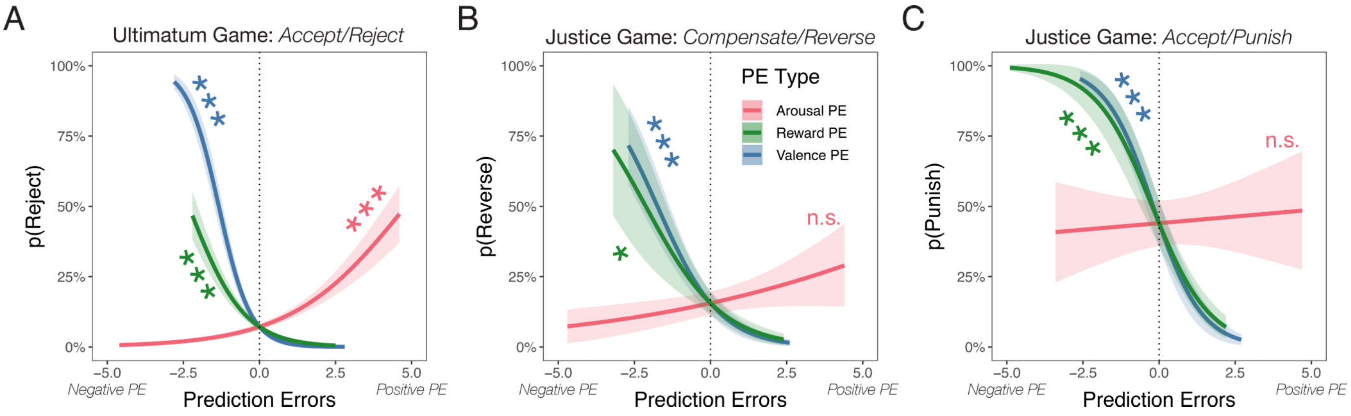

Figure 2. Emotion prediction errors underpin punitive behavior in the A) Ultimatum Game and B-C) Justice Game.

Participants completed either the Ultimatum Game (UG) or Justice Game (JG). The lines on each graph reflect the probability of different choice pairs including **A**) rejecting vs accepting in the UG, **B**) reversing vs compensating in the JG, and **C**) punishing vs accepting in the JG. The color of each line indicates Reward (green), Arousal (red), and Valence (blue) prediction errors. Negative values reflect negative prediction errors, indicating less money (reward), less pleasantness (valence), and less arousal than expected. Shaded areas reflect ±1 standard errors. Stars reflect significance at p < .001***, p < .01**, p < .05*.

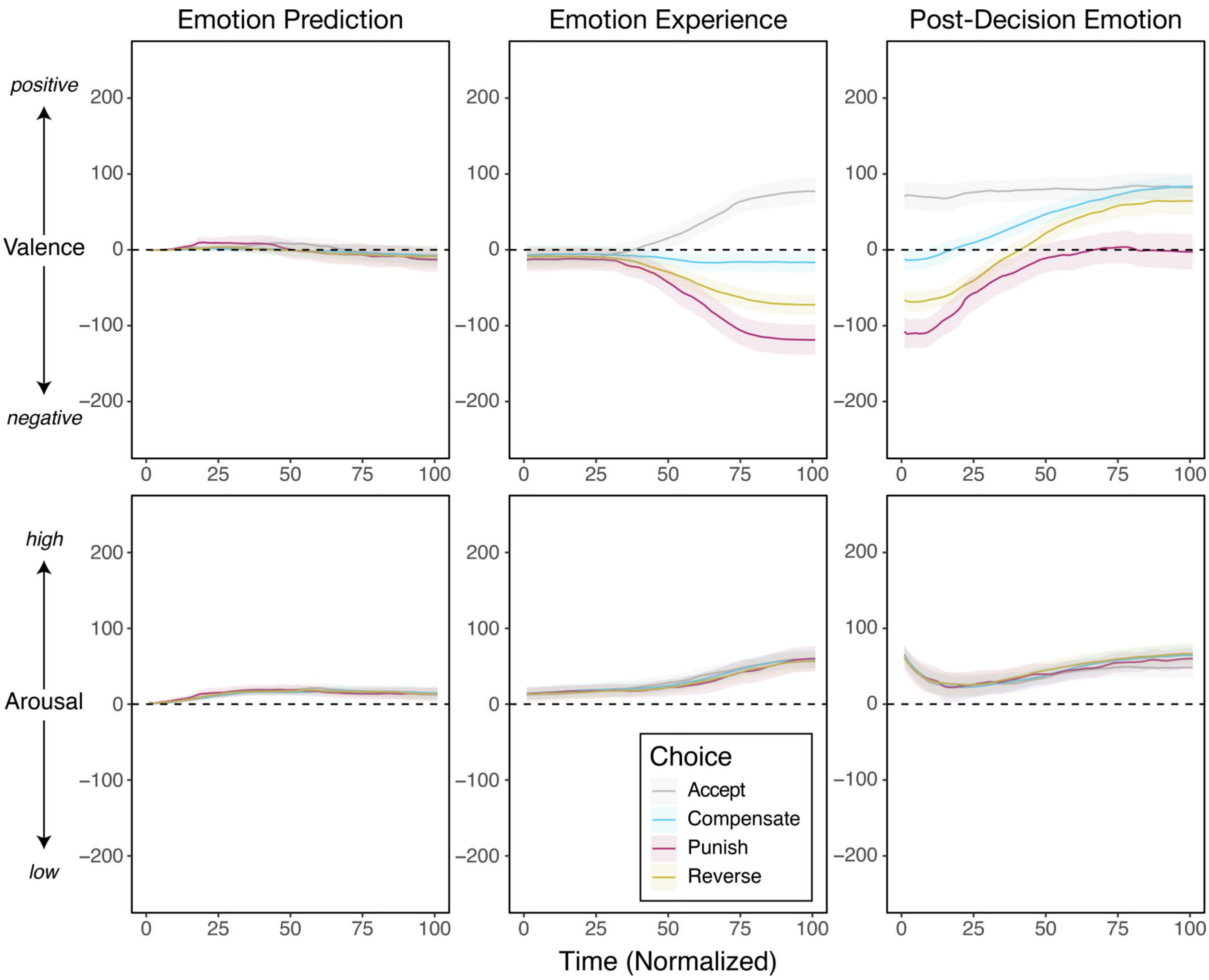

Figure 3. Temporal dynamics of emotional experiences during choice.

Participants used the dARM to continuously report their emotional experiences at every stage of the Justice Game. For all measurements, participants' data were normalized for the time it took to make a rating, such that 1 represents the start of the trial and 101 represents the final emotion rating. **Emotion prediction, experience, and post decision emotion**. The average valence and arousal are plotted over time. All shaded areas represent 95% CIs.

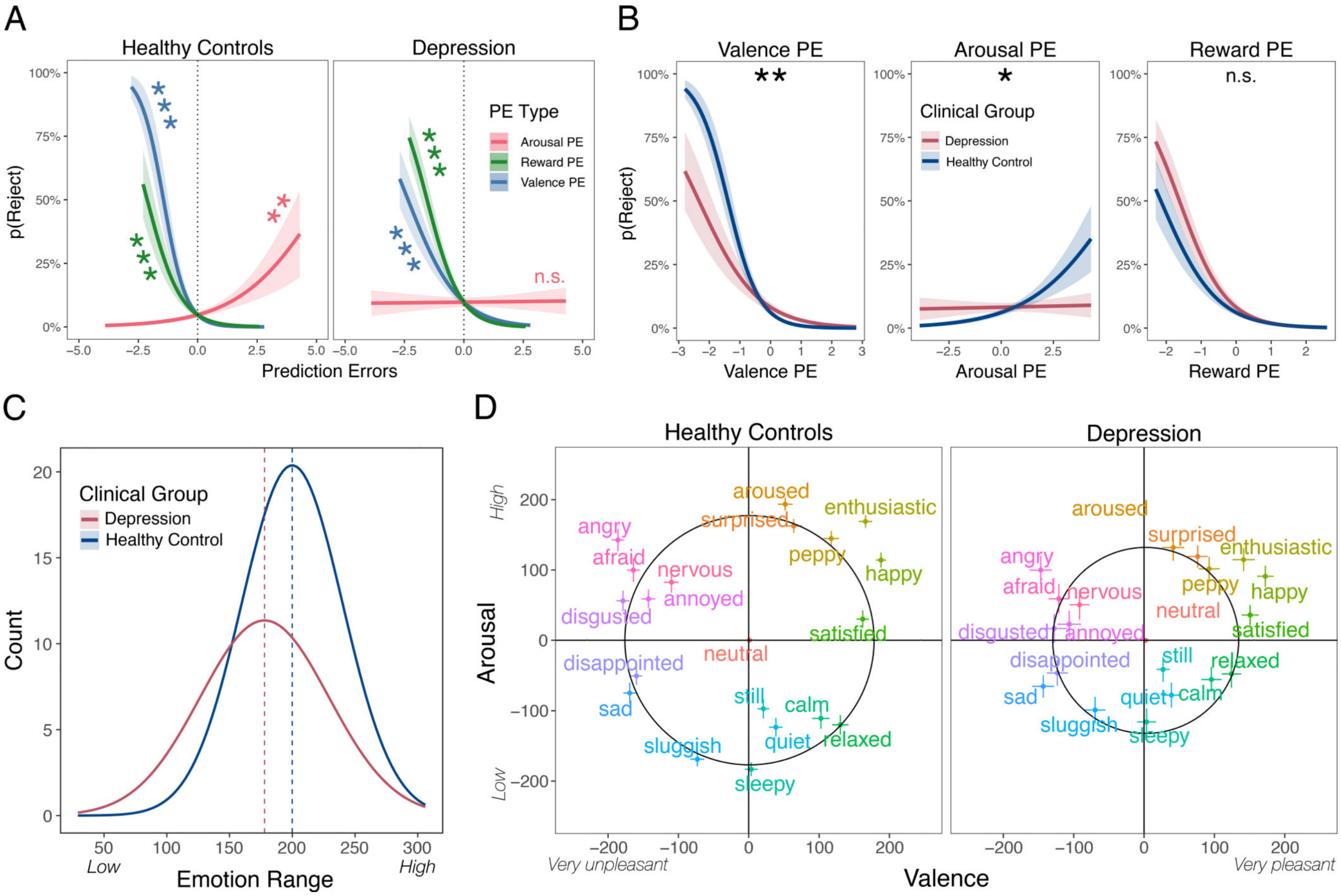

Figure 4. Experiment 4 Results.

A) Prediction error use in healthy controls and those at risk for depression. The probability of Rejecting the offer is plotted for all three types of prediction errors: reward, arousal, and valence for healthy controls (left) and those reporting significant levels of depression (right). Negative values reflect negative prediction errors, while positive values reflect positive prediction errors. Analyses represent separate regression effects. Shaded areas reflect ±1 SE. **B) Prediction errors plotted by group.** The use of each prediction error is plotted for both healthy controls and individuals at risk of depression. Analyses represent interactions between each prediction error and group. **C) Emotion range.** Each participant's emotion ratings were used to calculate the average distance of their ratings from neutral (i.e., the radius of their unique circumplex), thereby indexing their emotional range. Histograms represent normal distributions for both groups and dashed lines indicate their respective means. **D) Group-level average emotion ratings.** Participants rated 20 typical emotions using the dARM and the circles represent emotion range fits for each group. Error bars represent 95% CIs. Shaded areas reflect ±1 SE. Stars reflect significance at p < .001***, p < .01**, p < .05*.

Table 1.

Experiment 1: Valence prediction errors predict decisions to punish better than reward prediction errors

$\mathit{Punish}_{i,t} \sim \beta_0 + \beta_1\mathit{Reward\ PE}_{i,t} + \beta_2\mathit{Valence\ PE}_{i,t} + \beta_3\mathit{Arousal\ PE}_{i,t} + \varepsilon$

Variable	**Estimate (SE)**	**z**	**p**
Punish			
Intercept	−2.56 (0.18)	−14.26	$<.001^{***}$
Reward PE	−1.10 (0.14)	−7.62	$<.001^{***}$
Valence PE	−1.92 (0.16)	−11.75	$<.001^{***}$
Arousal PE	0.53 (0.09)	5.65	$<.001^{***}$

Note. Reward PEs are calculated by taking the difference between the experienced and predicted reward. Valence PEs and Arousal PEs are calculated by taking the difference between the experienced and predicted emotion. All variables were scaled but not mean-centered, as the 0 point on each scale refers to the meaningful instance where participants' expectations matched their experience. The model includes subject-specific random intercepts and slopes for Reward PE, Valence PE, and Arousal PE. The dataset includes 7,280 observations from 364 participants.

***p <.001.

Table 2.

Experiment 4: Individuals at risk of depression are less sensitive to unfairness

$\boldsymbol{Punish_{i,t} \sim \beta_0 + \beta_1 Depression_i + \beta_2 Unfairness_{i,t} + \beta_3 (Depression \times Unfairness)_{i,t} + \varepsilon}$

Variable	**Estimate (SE)**	**z**	**p**
Punish			
Intercept	−4.57 (0.48)	−9.56	<.001 ***
Depressed	1.42 (0.60)	2.38	.018 *
Unfairness	5.57 (0.38)	14.53	<.001 ***
Depressed × Unfairness	−1.80 (0.42)	−4.24	<.001 ***

Note. Unfairness is scaled and mean-centered. Depression is a binary variable with healthy controls (0) and those at risk of depression (1). The model includes subject-specific random intercepts and slopes for Unfairness. The dataset includes 7,020 observations from 351 participants.

*p < .05.

***p <.001.

Table 3.

Experiment 4: Individuals at risk of depression have selective impairment in emotion (but not reward) prediction errors

$$\boldsymbol{Punish_{i,t} \sim \beta_0 + \beta_1 rPE_{i,t} + \beta_2 vPE_{i,t} + \beta_3 aPE_{i,t} + \beta_4 Depression_i + \beta_5 (rPE \times Depression)_{i,t} + \beta_6 (vPE \times Depression)_{i,t} + \beta_7 (aPE \times Depression)_{i,t} + \varepsilon}$$

Variable	**Estimate (SE)**	**z**	**p**
Punish			
Intercept	−2.73 (0.22)	−12.58	<.001 ***
Reward PE	−1.27 (0.19)	−6.54	<.001 ***
Valence PE	−1.96 (0.22)	−9.03	<.001 ***
Arousal PE	0.49 (0.14)	3.58	<.001 ***
Depression	0.31 (0.30)	1.04	.296
Reward PE × Depression	−0.22 (0.25)	−0.88	.378
Valence PE × Depression	0.93 (0.29)	3.24	.001 **
Arousal PE × Depression	−0.47 (0.18)	−2.54	.011 *

Note. Reward PEs are calculated by taking the difference between the experienced and predicted reward. Valence PEs and Arousal PEs are calculated by taking the difference between the experienced and predicted emotion. All variables were scaled but not mean-centered, as the 0 point on each scale refers to the meaningful instance where expectations match experiences. Depression is a binary variable with healthy controls (0) and those at risk of depression (1). The model includes subject-specific random intercepts and slopes for Reward PE, Valence PE, and Arousal PE. The dataset includes 7,020 observations from 351 participants.

*p < .05.

**p < .01.

***p <.001.