Supplementary Note

Environmental Volatility and Learning Rate Estimation

In Simulation 6 in the main text, we simulated the PRO model in a 2-arm bandit task similar to a previously reported study. A reinforcement learning model was then fit to the trial-by-trial choice behavior of the PRO model in order to recover effective learning rates in stable and volatile periods. The reinforcement learning model is described by a learning law that tracks the value (V) of choices \( i \), and an actor component that determines the probability of making a particular response. To learn the value of each choice \( i \), we used a delta-learning rule:

\[
V_{i,t+1} = V_{i,t} + \alpha (R_{i,t} - V_{i,t})
\]

(1)

where \( R_{i,t} \) is the level of reward (0 or 1) observed for choice \( i \) on trial \( t \) and \( \alpha \) is a learning rate parameter. The probability of selecting a choice \( i \) was computed by a softmax function:

\[
P_i = \frac{e^{\gamma V_i}}{\sum e^{\gamma V}}
\]

(2)

where \( \gamma \) is a scaling parameter which determines the confidence in choice \( i \). The reinforcement learning model contained 5 free parameters: \( \gamma \), and four learning rate parameters \( \alpha \), one for each period in the task (Training, Volatile 1, Volatile 2, Stable). The estimated learning rates for the PRO model are shown in Fig. 4b in the main text, along with learning rates estimate for a lesioned version of the model in which surprise signals had no effect on learning.

Similarly, we implemented a Bayesian learner similar to one previously described that tracks reward probabilities and estimated environmental volatility. The Bayesian learner was trained using choice data generated by the PRO model, and for each period in the task, the mean estimated volatility was calculated (Fig. 4b in the main text).

Supplementary Discussion

Comparison with other models of performance monitoring

The PRO model suggests that error effects in mPFC derive essentially from a discrepancy (subtraction) between actual and expected outcomes. Theories of mPFC function depend heavily on the apparent role of mPFC in detecting and processing errors. Beginning with early ERP studies, effects of error have been routinely observed in human EEG and imaging studies. Theoretical accounts of error effects can be divided into several categories. One view treats mPFC as dealing with error qua error: error is an explicit
quantity which is signaled (if not calculated) directly by mPFC. An alternative view is that error is an *implicit* term which emerges from computational processes which do not directly calculate error.

**Explicit Calculation of Error**

The explicit view of error processing in turn leads to the question of what constitutes an error. That is, what is the computational form of the error calculation, and over what terms is this computation conducted? The notion of error as a discrepancy suggests a simple form for calculating error:

$$\text{Error} = \text{Expected} - \text{Actual}$$

This definition leaves open the questions of what quantity is expected, and what actual quantity is experienced. One possibility is that aversive events are the result of incorrect or inappropriate actions, and that the error computation compares *intended* actions to *actual* actions. Nonetheless, others have shown that error feedback leads to an ERN-like signal, even when the action was generated as intended. This suggests a comparison between actual and intended *outcomes*, consistent with neurophysiological findings in monkeys.

The PRO model builds on these accounts by specifying that the “expected” quantity reflects the conjunction of responses and outcomes. Furthermore, rather than reflecting *intended* response-outcome conjunctions, the PRO model treats the “expected” quantity as a more general prediction of the likelihood these conjunctions will occur, regardless of affective valence and whether they are intended or not. In this context, then, the PRO model casts error in a more general frame. Rather than reflecting differences between desired and actual outcomes or responses, error reflects how well or poorly future events are predicted, with mPFC strongly signaling surprising non-occurrences of predicted events, as well as the surprising occurrence of unexpected events.

**Implicit Error Signals**

Implicit calculation of error is perhaps best embodied by the conflict theory of mPFC. Under this view, mPFC signals response conflict, calculated as the product of the activation of mutually incompatible responses. Error is not directly calculated as a discrepancy, but is implicitly signaled by the continued, simultaneous activation of potential responses following the generation of an erroneous response. The logic is that when an error is committed in the presence of conflict, then the correct response is also likely to have been prepared, even though it was overwhelmed by the incorrect response process. Thus a state of conflict exists between the incorrect and correct response representations on error trials, whereas no such state exists on correct trials.

The conflict account of mPFC function is appealing due to the number of observed phenomena which it describes using the straightforward principle of behavioral conflict. Like the PRO model, conflict theory accounts for commonly observed effects in mPFC, including error and conflict, and the amplitude of the N2 as a function of accuracy. Given the array of effects in common which are described by the conflict and PRO models, it is an important question as to whether they make distinct predictions.

One data set that may discriminate between the conflict and PRO models focuses on partial errors. Burle and colleagues investigated the conflict theory using the Eriksen flanker task. Specifically, they looked at the amplitude of the ERN following partial errors, in which an incorrect response is prepared to
a certain extent but then suppressed (as indicated by sub-threshold electromyographic activity). While the conflict model predicts that ERN amplitude should decrease with time due to increased temporal separation between incorrect and correct responses, the authors found that ERN amplitude actually increased with time. This finding presents a challenge to the conflict account. The PRO model may account for this effect. In the main text, Simulation 3 (Fig. 2c) treats a slightly different situation in which the N2 amplitude in a flanker task is binned by RT. The human data show a positive correlation between N2 amplitude and response time. The PRO model accounts for this positive correlation. Specifically, mPFC activity increases with unexpectedly delayed outcomes, because in the PRO model, activity predicting an expected outcome continues to rise unchecked until an actual outcome occurs to meet (suppress) the prediction. Thus, longer delays until the action or feedback correlate with greater mPFC activity. Variation in response generation due to processing noise in response units in the PRO model could be a source of delay; slower-than-average responses following partial errors (or any other expected action/outcome) would therefore be expected to result in increased model activity.

Reinforcement Learning

The PRO model bears a close resemblance to models which suggest that mPFC activity reflects a temporal difference error. Like these previous models, the PRO model implements a reinforcement learning algorithm based on temporal difference learning 5, 14. However, there are two key differences introduced by the PRO model. First, the PRO model learns to predict multiple outcomes using a vector reinforcement term which indicates the occurrence of one or more events. In contrast, previous models 5, 14 use a scalar reinforcement learning signal which does not signal the occurrence of an event as such, but instead reports the valence of an event (i.e., rewarding or aversive). As noted in the main text, the reinforcement signal in the PRO model does not ascribe a particular valence to the observed event; unexpected aversive outcomes are learned in the same manner as unexpected rewarding events. Previous reinforcement learning models of mPFC, however, incorporate affective valence as a part of the learning signal: worse-than-expected occurrences are assigned negative values while better-than-expected occurrences are given positive values. In the case of composite events which may have a rewarding component as well as an aversive component, the scalar learning signal reports some combination of the rewarding and aversive components of the composite event.

The significance of these differences is that while the PRO model independently represents predictions of each possible future event, previous models based on reinforcement learning represent combinations of the value of future events. In these previous models, the ERN observed in the mPFC is modeled as the (negative) TD error. Consequently, these models are able to signal errors but not unexpected affectively positive events, such as a surprising win of a gamble with low probability of winning. This is a significant limitation, given recent findings that mPFC shows such effects 7, which the PRO model is able to simulate (Fig. 4c, main text).

References


Figure S1

Nature Neuroscience: doi:10.1038/nn.2921
**Figure S1. PRO model diagram.** The PRO model consists of three components. The representation component (blue) learns to represent the probability of various possible combinations of responses and outcomes (“response-outcome conjunctions”) depending on the incoming stimuli. The predicted probabilities serve as a basis for the control signal to the actor component of the model (green), which maps stimuli to actions. Weights from the representation to actor components are adjusted by a gating signal which indicates the affective valence of an event, i.e. good or bad. The critic component (red) implements a variant of temporal difference learning, but with multiple predictions instead of a single prediction. Specifically, the learning signal is computed as the difference between an actual outcome (i.e. response and outcome conjunction, whether good or bad) and the predicted outcome, based on incoming stimulus signals. Decomposed into positive and negative surprise signals, $\omega^P$ and $\omega^N$, the learning signal is used to modulate the rate at which associations between task stimuli and response-outcome conjunctions are learned. Here we use the term actor to refer to the mechanisms that map stimuli to responses, and not to refer to the unit that predicts the outcome of actions, even though the latter could be thought of as a “cognitive actor” to the extent it generates predictions and is computationally similar to the actor in previous actor-critic models. See text for description of model terms.