Title: Medial prefrontal cortex as an action-outcome predictor

Authors: William Alexander & Joshua Brown

Supplementary Methods

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[1]. 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:

$$V_{i,t+1} = V_{i,t} + \alpha(R_{i,t} - V_{i,t}) \tag{1}$$

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

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

where γ is a scaling parameter which determines the confidence in choice *i*. The reinforcement learning model contained 5 free parameters: γ, and four learning rate parameters α , 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[1] 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 [2, 3], 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 [4]. Nonetheless, others have shown that error feedback leads to an ERN-like signal, even when the action was generated as intended [5]. This suggests a comparison between actual and intended *outcomes*, consistent with neurophysiological findings in monkeys [6].

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 [7]. 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 [8-10]. 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 [9], and the amplitude of the N2 as a function of accuracy [11]. 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 [12] investigated the conflict theory using the Eriksen flanker task [13]. 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, similar to a positive correlation found between RT and BOLD activity in human mPFC[14, 15]. 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.

The PRO model also suggests a reconciliation between conflict and error likelihood theories of mPFC function. Recent work (Fig. 2b, right) has argued that mPFC effects are consistent with conflict but not error likelihood effects[11]. The N2 is greater for slower and more accurate trials than faster trials with error likelihood. On its face, this seems to contradict the finding of error likelihood effects in mPFC[16]. Nevertheless, the PRO model shows error likelihood effects. How can this be? The answer lies in the distinction between *cue-based error likelihood* and *RT-based error likelihood*. The PRO model shows greater activity when a cue predicting a higher error likelihood appears relative to a lower error likelihood cue (Fig. 1b). This cue-based error likelihood signal is averaged over the entire RT distribution for each corresponding cue. If instead a single cue is isolated and the responses analyzed by RT, then those trials in slower RT bins will yield greater activation with a lower error likelihood, and those cues in the faster RT bins will yield less activation with greater error likelihood (Fig. 2b, right). Overall, the guiding principle is that error likelihood effects are found across cued conditions (cue-based error likelihood), but inverted error likelihood effects are found across the speed-accuracy tradeoff curves of a single condition (RT-based error likelihood).

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, 17]. 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, 17] 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

1. Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. Learning the value of information in an uncertain world. *Nat Neurosci* **10**, 1214-1221 (2007).
2. Gehring, W.J., Coles, M.G.H., Meyer, D.E. & Donchin, E. The error-related negativity: An event-related potential accompanying errors. *Psychophysiology* **27**, S34 (1990).
3. Falkenstein, M., Hohnsbein, J., Hoorman, J. & Blanke, L. Effects of crossmodal divided attention on late ERP components: II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology* **78**, 447-455 (1991).
4. Scheffers, M.K. & Coles, M.G. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *J Exp Psychol Hum Percept Perform* **26**, 141-151 (2000).
5. Holroyd, C.B. & Coles, M.G. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psych. Rev.* **109**, 679-709 (2002).
6. Ito, S., Stuphorn, V., Brown, J. & Schall, J.D. Performance Monitoring by Anterior Cingulate Cortex During Saccade Countermanding. *Science* **302**, 120-122 (2003).
7. Jessup, R.K., Busemeyer, J.R. & Brown, J.W. Error effects in anterior cingulate cortex reverse when error likelihood is high. *J. Neurosci.* **30**, 3467-3472 (2010).
8. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S. & Cohen, J.C. Conflict monitoring and cognitive control. *Psychological Review* **108**, 624-652 (2001).
9. Yeung, N., Cohen, J.D. & Botvinick, M.M. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev* **111**, 931-959 (2004).
10. Jones, A.D., Cho, R., Nystrom, L.E., Cohen, J.D. & Braver, T.S. A computational model of anterior cingulate function in speeded response tasks: Effects of frequency, sequence, and conflict. *Cog Aff Behav Neurosci.* **2**, 300-317 (2002).
11. Yeung, N. & Nieuwenhuis, S. Dissociating response conflict and error likelihood in anterior cingulate cortex. *J Neurosci* **29**, 14506-14510 (2009).
12. Burle, B., Roger, C., Allain, S., Vidal, F. & Hasbroucq, T. Error negativity does not reflect conflict: a reappraisal of conflict monitoring and anterior cingulate cortex activity. *J Cogn Neurosci* **20**, 1637-1655 (2008).
13. Eriksen, B.A. & Eriksen, C.W. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* **16**, 143-149 (1974).

14. Carp, J., Kim, K., Taylor, S.F., Fitzgerald, K.D. & Weissman, D.H. Conditional Differences in Mean Reaction Time Explain Effects of Response Congruency, but not Accuracy, on Posterior Medial Frontal Cortex Activity. *Front Hum Neurosci* **4**, 231 (2010).
15. Grinband, J.*, et al.* The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *Neuroimage* **57**, 303-311 (2011).
16. Brown, J.W. & Braver, T.S. Learned Predictions of Error Likelihood in the Anterior Cingulate Cortex. *Science* **307**, 1118-1121 (2005).
17. Holroyd, C.B., Yeung, N., Coles, M.G. & Cohen, J.D. A mechanism for error detection in speeded response time tasks. *J Exp Psychol Gen* **134**, 163-191 (2005).

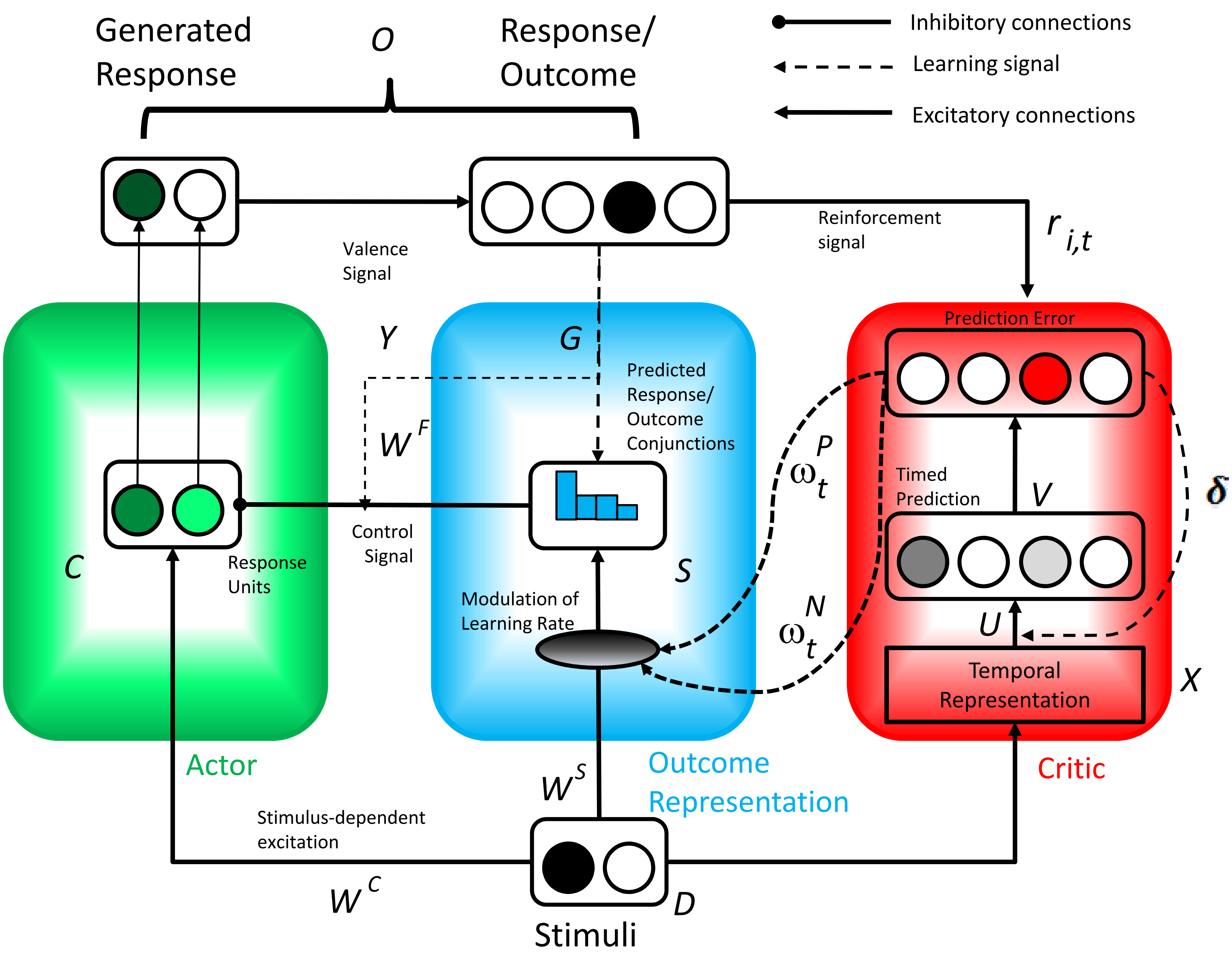
Generated Response
O
Response/ Outcome
Inhibitory connections
Learning signal
Excitatory connections
Valence Signal
Reinforcement signal
$r_{i,t}$
Y
G
Predicted Response/ Outcome Conjunctions
Prediction Error
W^F
ω^P_t
Timed Prediction
V
δ
C
Control Signal
Response Units
S
Modulation of Learning Rate
ω^N_t
U
Temporal Representation
X
Actor
Outcome Representation
Critic
W^S
Stimulus-dependent excitation
W^C
D
Stimuli

Figure S1

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, ω^P and ω^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.