Model-free and model-based reward prediction errors in EEG

Thomas D. Sambrook[1]

Andy J. Wills[2]

Ben Hardwick[2]

Jeremy Goslin[2]

[1]School of Psychology, University of East Anglia

[2]Cognition Institute, School of Psychology, University of Plymouth

Abstract

Learning theorists posit two reinforcement learning systems: model-free and model-based. Model-based learning incorporates knowledge about structure and contingencies in the world to assign candidate actions with an expected value. Model-free learning is ignorant of the world's structure; instead, actions hold a value based on prior reinforcement, with this value updated by expectancy violation in the form of a reward prediction error. Because they use such different learning mechanisms, it has been previously assumed that model-based and model-free learning are computationally dissociated in the brain. However, recent fMRI evidence suggests that the brain may compute reward prediction errors to both model-free and model-based estimates of value, signalling the possibility that these systems interact. Because of its poor temporal resolution, fMRI risks confounding reward prediction errors with other feedback-related neural activity. In the present study, EEG was used to show the presence of both model-based and model-free reward prediction errors and their place in a temporal sequence of events including state prediction errors and action value updates. This demonstration of model-based prediction errors questions a long-held assumption that model-free and model-based learning are dissociated in the brain.

Introduction

Choice behaviour becomes amenable to explanation if it is assumed that *value* is a construct that is internally represented. The construct of value is what unites explanations of decision-making and reinforcement learning insofar as it is the different values of candidate actions that determine what decisions are taken, and the revision of action values in the light of feedback that constitutes learning. It is also hoped that value may come to be materially

defined within the brain, thus uniting behavioural, neural and subjective explanations of choice (Glimcher 2009).

Its attractiveness as a unitary construct notwithstanding, there appear to be multiple forms of action valuation. A particularly important distinction lies between goal-directed and habitual action. Goal-directed action is led by the value of the outcome it hopes to attain, while habit-led action occurs because the mere action itself has acquired value owing to previous reward. In terms of associative links, goal-directed and habitual actions arise from action–outcome and stimulus–response associations respectively (Dickinson 1985). This has wide ramifications. When outcomes are known to be unattainable or have ceased to have value (e.g. food for a sated animal), goal-directed decision making can make use of the action-outcome link to accordingly down-value the action. This is not possible in habitual behaviour because there is no representation of the outcome. Instead, the behaviour is under the control of the stimulus and so is selected inappropriately. The consequences of inappropriate habitual behaviour can range from the inconvenient (taking the turning to work on a Saturday when the goal was a shopping trip) to the grave (perseveration of maladaptive cycles of harmful behaviour such as substance abuse). Nevertheless, in a stable environment, the habit system is more computationally efficient than the goal-directed system (de Wit and Dickinson 2009), and an adaptive agent working under cognitive constraints should incur benefits from delegating a large portion of action control to the habit system.

A dual process theory such as this naturally accommodates wide behavioural variation, and so the habit-led / goal-directed distinction, and the representation of outcomes or their neglect, provides a powerful overall framework for explaining both inter- and intra-individual differences in decision making (de Wit, et al. 2012; Eppinger, et al. 2013). Further progress requires us to understand the mechanism by which goal-directed and habitual systems assign value to actions. Here, computational models of learning can be usefully

brought to bear insofar as they signpost optimal information processing (to which natural selection tends, if not unerringly), express key computational terms for which we may hope to find behavioural or neural correlates, and express this information in unequivocal terms (Wills, et al. 2017; Wills and Pothos 2012). The properties of goal-directed and habit led *behaviour* strongly suggest they are underlain by two distinct forms of *learning*, termed model-based and model-free (Daw and O' Doherty 2013). Model-based learning maintains separate representations of the likelihood with which an action will lead to an outcome, and that outcome's incentive value. The value of an action ultimately rests on the product of these two terms, the classic formulation of expected value derived by Pascal (Hald 2003). Because of this explicit representation of the contingency x incentive value structure that underlies an action value, an agent can prospectively change action values when either the likelihood of the outcome or its incentive value is changed. In practice, multiple sub-states are likely to lie between an initial reward directed action and its consumption, necessitating the representation of an extensive network of transition probabilities. In model-based learning an animal thus models the world, albeit strictly in terms of its ability to afford reward. In contrast, model-free learning represents neither the likelihood of outcomes nor their incentive value, nor indeed anything else. Deprived of the constituent terms required to generate expected value, it relies on a different mechanism of action valuation, which is to simply maintain a record of how much reward the action has previously incurred, using this retrospective valuation as a proxy for expected value. The advantage of such a mechanism is that it is computationally cheap.

Computationally, model-based and model-free learning contain terms that are in some cases shared and in other cases unique to themselves. Pre-eminent in model-free learning is the reward prediction error (RPE), the difference between expected and obtained reward, which is used to adjust action values upwards or downwards (Bush and Mosteller 1955;

Rescorla and Wagner 1972; Sutton and Barto 1998). RPEs are not used to update action values in model-based learning. A quite separate error term is generated, the state prediction error (SPE), which reflects the degree to which an outcome, or intermediate state, was unexpected, and this error term is used to adjust the network of action-outcome contingencies (Glascher, et al. 2010). A term that both model-free and model-based learning share is the action value. However, for reasons described earlier, the two systems will often produce different values for an action.

By comparing participants' behaviour in a reinforcement learning task to a simulation of model-based and model-free learning it is possible to assess the evidence for which of the two forms of learning are controlling behaviour. Furthermore, by correlating the values of the model-free and model-based computational terms described above with neural activity, it should also be possible to identify the neural correlates of these two kinds of learning. This provides a useful counterpart, drawn from the normal population, to lesion data, which has previously been the chief source for inferring the neural instantiation of the two learning variants. The computational approach has been used to show model-free RPEs in the striatum (O'Doherty, et al. 2004; Seymour, et al. 2012) and an SPE in the dorsolateral prefrontal cortex (Glascher et al., 2012). This pattern of results corresponds to lesion work suggesting dissociated model-free striatal and model-based prefrontal cortex systems (Killcross and Coutureau 2003; Yin, et al. 2004; Yin, et al. 2005).

Neural correlates also comprise evidence that computational models accurately describe the process of choice and learning. In the case where the neural correlate of a computational term cannot be found, this may be ascribed to methodological constraints of brain imaging. In contrast, where evidence is found for a computational term not predicted by the models, this is strong evidence that those models are incomplete. Daw, et al. (2011), for example, in a study that forms the basis for the present one, provided fMRI evidence that, on

arrival of feedback following an action, the brain computes a RPE to both model-free and model-based action values. This is despite the fact that model-based learning does not use RPEs since these embody a mechanism of incremental adjustment of prior value which model-based learning eschews in favour of prospective estimates of value based on what is known about the world. Daw et al.'s demonstration suggests that the sharp distinction between model-based and model-free learning described by computational models may not be so neatly realised in the brain, and that new hybrid models may be needed. One possibility is that goal-directed action values are not, as has been supposed, built exclusively from prospective evaluation of outcome value and likelihood, but are tempered to some degree by a record of previous reward obtained relative to (model-based) expectation. Alternatively, a model-based prediction error may serve no function for model-based learning but instead may be used as additional input to model-free learning.

The present study used EEG to assess the evidence for model-based prediction errors, acting as a counterpart to Daw et al.'s (2011) fMRI study. The importance of complementary EEG and fMRI evidence, given these methods' respective paucity in spatial and temporal resolution is well known. In the present case, the temporal precision of EEG may bring the greater gains however. The computational terms by which model-free and model-based learning can be distinguished may not be spatially segregated at the resolution attainable by fMRI. In contrast, computations associated with the prediction error mechanism, whether it is model-based or model-free, must respect a basic ordering, namely action valuation, action, outcome receipt, prediction error generation and action valuation update. We thus have a prior basis for hoping for temporal discrimination, albeit within the limits imposed by the ERP technique, and the further possibility that these steps reflect a cascade rather than discrete events. One of the goals of the present study is to use EEG's temporal resolution to

demonstrate that model-based prediction error effects are not merely residual effects of model-based action values, either prior or subsequent to action.

This study used the two step task of Daw, et al. (2011) which affords both model-based and model-free learning strategies. A computational model containing both model-free and model-based components was used to model participants' behaviour, with free parameters fitted on a participant-wise basis including a parameter determining the relative influence of the two components. The trial by trial values of the computational terms in the model were used as regressors in a multiple regression against scalp voltage to establish the neural correlates of each term while controlling for the others. Four such terms were investigated. We expected to find a model-free RPE, since this component, described variously as the "feedback related negativity" (Holroyd and Coles 2002) or "reward positivity" (Holroyd, et al. 2008) is routinely observed in model-free learning tasks. We also expected to find a neural correlate of SPE insofar as unexpected transitions between states are salient events. We also assessed the evidence for neural correlates of action values since these are, neuroeconomically speaking, the critical term for decision making inasmuch as they predict choice. Finally, we assessed evidence for a model-based RPE: the central purpose of this study.

Methods

Participants

The study was approved by the ethics committee of the Faculty of Health and Human Sciences at the University of Plymouth. Sixty five female students of the University of

Plymouth participated for course credit and an opportunity to win money. All participants were under 29 years, had no history of neurological damage or other significant health problems, and were not on medication at the time of the experiment. No other information was recorded. Four participants were excluded for excessive artefacts (>75% trials lost). Sixteen participants were excluded on the basis of failure to learn or of inadequate task attention (see Results section) leaving a final sample of forty five participants.

Task rationale

The task was a variant of Daw et al.'s (2011) two-step task as used by Gillan, et al. (2015). The probabilistic structure of the task is shown in Fig 1a. Participants chose from one of two fractals at a starting state (state 1) and were taken to one of two possible intermediate states (state 2) at which they were shown a single fractal. Each state 1 fractal had both a likely (70%) and unlikely (30%) following state 2, with these probabilities explicitly known to the participant. Following state 2, the participant was then shown the outcome, either reward or no reward. The probabilities with which each state 2 led to reward were initialised at hidden values between .25 and .75 and allowed to drift slowly and independently over time within those boundaries. Participants could thus use the observed delivery of reward or its omission to estimate these hidden probabilities at any given time. It should be stressed that reward was contingent entirely on the state 2 reached. Over time, based on outcomes, participants should be expected to differentially value the two state 2's and, insofar as choice at state 1 determines to some degree the state 2 reached, should also differentially value the two state 1 choices available. These valuations, both of state 1 choices and state 2's reached, were modelled on a trial by trial basis by a computational model which contained both model-based and model-free components. The key design feature of the task is that these

components will come to ascribe different values. The fit of each component's output to behaviour thus indicates the degree to which each is in control. More importantly for the present paper, fitting each component's output to neural data can be used to reveal where and when model-based and model-free process occur in the brain. The reason why model-free and model-based valuation diverges in this task is best illustrated by considering the case of a state 2 which is arrived at via an unlikely (30%) transition and then produces a reward. For both model-based and model-free components this raises the value of that state 2. However, regarding the state 1 choices, the model-based component will raise the value of the *unchosen* option since its model of the task incorporates the knowledge that this choice is more likely to result in transition to the recently rewarded state 2. The model-free component does not represent this information and will simply raise the value of the state 1 choice selected since it led to reward on that trial. While this is a specific scenario, it is generally the case that the neglect or representation of transitions will result in different model-free and model-based valuations throughout the task.

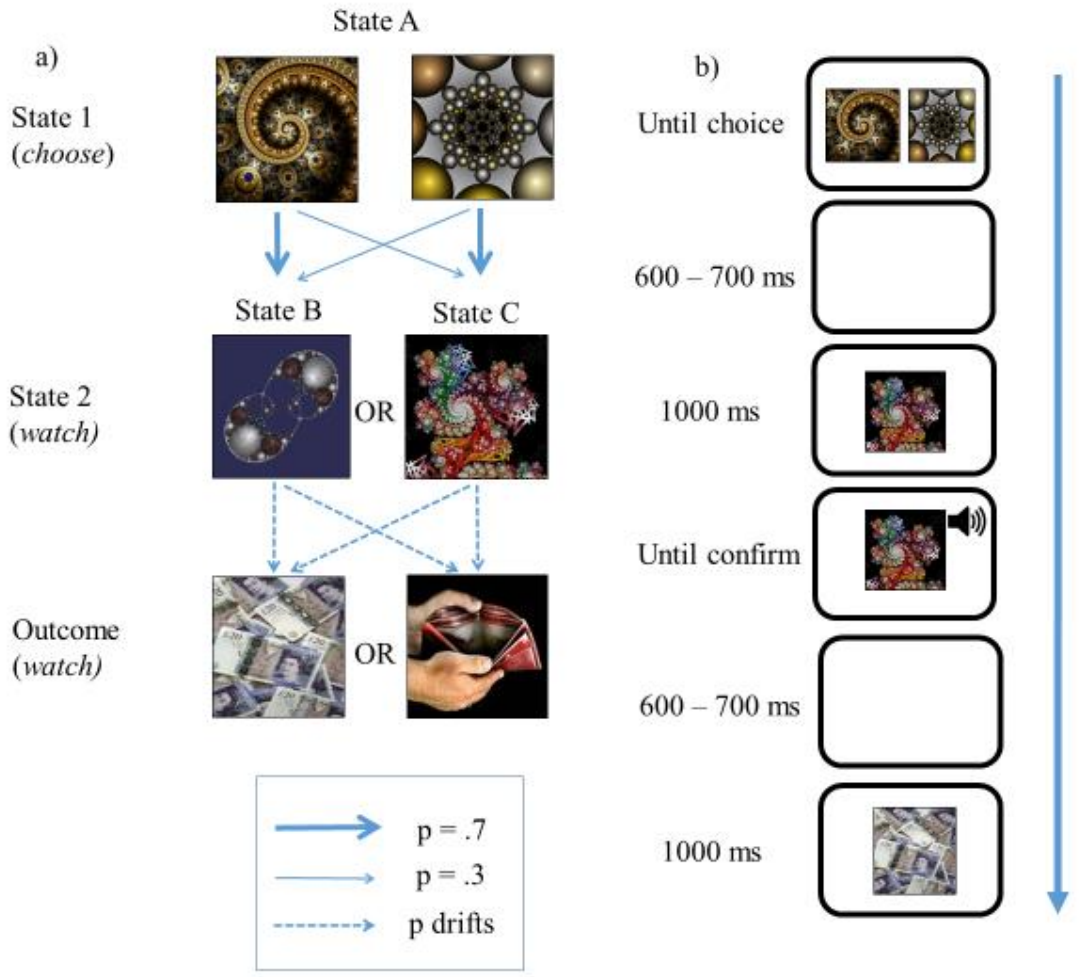

Figure 1. a Structure of the two step task b Sequence of events on one trial

Task implementation and execution

The experimental task was presented using E-Prime software. Fractals were taken from http://www.fractalsciencekit.com. Four fractals were used to denote the two choices available at state 1 and each of the two state 2's, with these randomly allocated for each participant. Participants were given the requisite knowledge to engage in model-based learning. That is, they were shown the state 1 fractal associated with each state 2, informed of the 70% transition probability and alerted to the need to track the probability with which each state 2 led to reward. Participants undertook blocks of sixty trials with a short break between blocks.

Because trials were self-paced, the number of blocks varied (M = 12.29, SD = 1.67). Each trial followed the format shown in Fig 1b. Participants were presented with a choice of two state 1 fractals and chose one via a response box. A blank screen was presented, followed 600 – 700 ms later by one of the state 2s. Participants then waited 1000 ms until a tone sounded, at which time they made a confirmatory key press on the response box to progress the trial. Early responses resulted in a penalty and these trials were not used in subsequent analysis. A further blank screen was shown, followed 600 – 700 ms later by the trial outcome. At the conclusion of a block, participants earned £0.30 if they had achieved thirty one or more reward outcomes. To assess task attention, participants were required at the end of each block to indicate which of the two state 2s they thought was most likely to lead to reward and which of the two state 1 fractals was the better choice.

Computational model

The computational model and associated parameter fitting procedures were implemented in R (R Core Team, 2017) using the slpMBMF function of the catlearn package (Catlearn Core Team, 2017). The computational model is based on that employed by Gillan, et al. (2015). It contains separate model-free and model-based components that work in parallel to derive their own estimates of the value of the fractals in the task. Then, to establish for each fractal a single overall value to be used for action selection, the model-based and model-free estimates are combined into an average that is weighted by the participant's bias towards model-free or model-based learning. This bias, along with further parameters, was estimated from the participant's behaviour.

Model-free component. The model-free component was SARSA(λ) temporal difference learning (Sutton and Barto 1998). This algorithm controls learning at state 1 and 2 of the task which, collectively, can take three forms, the opening state, denoted s_A which is always experienced, and two alternative state 2's, denoted s_B and s_C, only one of which is encountered on a given trial, *t*. At these states, an action, *a* is taken and at the trial's conclusion a reward, *r* is experienced. States and actions at state 1 and 2 are denoted as s_1, a_1 and s_2, a_2 and rewards that follow state 1 and 2 as r_1, r_2 (with the former always zero in the current task). The model-free value Q_{MF} for a given state action pair is updated using α, the learning rate and δ_{MF}, the model-free prediction error as follows

$$Q_{MF}(s_{i,t}, a_{i,t}) = Q_{MF}(s_{i,t}, a_{i,t}) + \alpha.\delta_{MF,i,t}$$

where

$$\delta_{MF,i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t})$$

This is the general form of the algorithm, but in the present task, note that actions are not performed at s_B and s_C (thus the state action values there are more formally described as a state values), that, as already noted, rewards are not incurred following s_1, so $\delta_{MF,1,t}$ is given entirely by the Q_{MF} value of the state 2 transitioned to, and that conversely no state is reached after state 2 and so $\delta_{MF,2}$ is determined only by the reward r_2. At the trial's conclusion, the eligibility parameter λ is used to modulate an additional stage-skipping update of the state 1 action by the state 2 prediction error

$$Q_{MF}(s_{1,t}, a_{1,t}) = Q_{MF}(s_{1,t}, a_{1,t}) + \alpha.\lambda.\delta_{MF,2,t}$$

Model-based component. Learning by the model-based component uses fixed knowledge regarding how the action taken at state 1 probabilistically determines the state 2 reached and combines this with that state 2 value to obtain action values for state 1.

$$Q_{MB}(s_A, a_j) = P(s_B|s_A, a_j)\, Q_{MF}(s_B) + P(s_C|s_A, a_j)\, Q_{MF}(s_C)$$

Two actions are available at state 1, pressing the left and right keys, denoted here as a_A and a_B. The transition probabilities, which were known to participants, were $P(s_B|s_A, a_A) = 0.7$; $P(s_C|s_A, a_B) = 0.7$; and $P(s_B|s_A, a_B) = 0.3$; $P(s_C|s_A, a_A) = 0.7$

Hybrid action value. For the purposes of action selection only, the state 1 Q values from the model-free and model-based components are combined into a hybrid, Q_H. This is a weighted average of the two individual values, Q_{MF} and Q_{MB}, with the weight given by ω, the degree to which the participant's observed behaviour was model-free or model-based. Thus

$$Q_H(s_A,a_j) = \omega.Q_{MB}(s_A,a_j) + (1- \omega).Q_{MF}(s_A,a_j)$$

This hybrid value is for action selection only and plays no role in learning. For learning, model-based and model-free components retain their separate Q values, updating these over trials as described above.

Parameter fitting

The parameters ω, α and λ were fitted on a participant-wise basis using maximum likelihood on observed choices. This necessitated the incorporation of a choice rule. A standard softmax rule was used, incorporating the inverse temperature parameter β (also fitted) to derive the probability, P, of each state 1 choice

$$P(a_{i,t} = a|s_{i,t}) = \exp[\beta.Q_H(s_{i,t}, a)] / \Sigma_{a'} \exp[\beta.Q_H(s_{i,t}, a)]$$

Each participant was fitted individually using the L-BFGS-B method (Byrd et al., 1995) of the *optim* function in R (R Core Team, 2017). In order to increase the stability of the neural regressors, and following Daw, et al. (2011), a second stage of fitting was conducted, where all participants were re-fit with all parameters except ω set at their median in the first-stage fit. Full source code for our fitting procedures will be published at: www.willslab.org.uk/ply116

Additional regressors

The analyses to come require two key regressors that do not feature in the computational model described above. One is the model-based prediction error δ_{MB}, the existence of which is not predicted by standard models, but for which we seek evidence. This is derived analogously to δ_{MF}, though using the difference of the model-based value of a state 1 action and the reward encountered at state 2 (always zero in this task) plus the model-free value of the state 2 transitioned to (since only model-free values are held at state 2).

$$\delta_{MB,i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MB}(s_{i,t}, a_{i,t})$$

The second is the SPE, which in model-based learning is used to update transition matrices. In keeping with Daw et al. (2011) and other authors using the two stage task, we do not attempt to model updating of transition matrices, assuming that the participant represents these as remaining fixed at .7/.3 throughout the task. Nevertheless, because unexpected transitions should still be expected to generate an SPE, this should be included as a regressor to control for its effect on the EEG. With a transition matrix based on two fixed values, SPEs can likewise only take two values, so the SPE regressor was entered as a dummy variable of zero or one for expected and unexpected transitions.

EEG recording

EEG data were collected from 61 Ag/AgCl active electrodes (actiCAP, Brain Products, Gilching, Germany) mounted on an elastic cap and arranged in a standard International 10–20 montage. Electrodes were referenced to the left mastoid and re-referenced off-line to the average of left and right mastoid activity. Vertical eye movement was monitored by a right suborbital electrode, and horizontal eye movement was monitored using an electrode on the right external canthus. Electrode impedances were kept below 20 kΩ. EEGs were amplified using a BrainAmp amplifier (Brain Products), continuously sampled at 500 Hz, filtered offline with notch filters at 60 Hz (screen refresh) and 50 Hz (AC interference) followed by a .1 Hz high pass filter and 30 Hz low pass filter and then down sampled to 125 Hz. Separate EEGs were created for activity following presentation of state 2 and outcomes. In each case, EEGs were time-locked to 200 ms before the onset of the feedback to 1000 ms afterward, and then were baseline-corrected using the interval −200 – 0 ms. Eye movement artefacts were

removed using a criterion of a voltage change exceeding 75 μv/200 ms in eye electrodes. Note that after the presentation of the initial choice (which is not analysed), all stimuli were presented in the centre of the screen, at fixation, thus there is no reason to suppose that saccades will occur, but more importantly, no means by which a confound of prediction error coding with stimulus position induced saccades could occur. Other non-specific artefacts were removed using a criterion of any electrode showing either a voltage change exceeding 40 μv/ms, a voltage value exceeding +/−200 μv relative to baseline, or activity across the epoch of below 2.5 μv. The percentage of trials retained was 79.8%. with a minimum of 63.1% (372 trials) for any one participant. Electrodes which malfunctioned in the course of an experiment were substituted using topographic interpolation (Perrin, et al. 1989).

Data Analysis

Regression models. Analyses were directed towards establishing evidence for neural representation of a model-based RPE, δ_{MB}, and so to achieve this, EEG voltage was regressed against the computational model's estimate of this term. Regressions were performed on the EEG following arrival of state 2 feedback since there was no possibility of model-based learning later in the trial. Regression models were performed on a univariate basis, i.e. an independent model calculated for each combination of the 61 electrodes and 150 sample points, and individually for each participant. The resulting beta values for regression coefficients of interest, foremost of these being δ_{MB}, were plotted at representative electrodes or as scalp maps (Figures 2, 4 and 5). To assess significance at the group level while controlling for the multiple comparisons resulting from mass univariate testing, the cluster randomisation procedure of Maris and Oostenveld (2007) was used. One sample t-tests (test

value = 0, N = 45) were performed on the beta values at each electrode / sample point and temporally and spatially contiguous significant points were grouped into a cluster. This was assigned a cluster-t value based on the sum of univariate t values and was then assigned a p value based on a comparison against a Monte Carlo generated distribution (10,000 iterations) of cluster-t values derived from a null distribution.

Regression models were conducted in the context of a high correlation between δ_{MB} and δ_{MF} (r = .72) and more moderate correlations between δ_{MB} and other terms likely to be represented in the EEG, the SPE (r = -.31) and the updated model-free Q value of the state 2 reached (r = .32). The effects of SPE and Q were controlled by including them as additional regressors. This was not possible in the case of δ_{MF} because of the collinearity arising from its strong correlation with δ_{MB}. Alternative methods were thus employed to avoid wrongly attributing the effects of δ_{MF} to δ_{MB}. As a preliminary test, we follow Daw, et al. (2011) who used a regressor of the form $\delta_{MB} - \delta_{MF}$ to describe the extent to which δ_{MB} differed from δ_{MF} on a given trial. By adding this term to a regression model already containing δ_{MF} they argued it was possible to describe the variance explained by δ_{MB} that was not already explained by δ_{MF}. While this difference term serves a useful means of establishing the existence of δ_{MB}, it is nevertheless a poor guide to the temporospatial profile of δ_{MB} itself rather than where its effect on scalp voltage is distinguishable from that of δ_{MF}. As Daw et al. note, modelling the residual effect of δ_{MB} on top of δ_{MF} rather than the effect of δ_{MF} on top of δ_{MB} merely reflects the theoretical status quo: that δ_{MF} is already assumed to be present. Consequently, the $\delta_{MB} - \delta_{MF}$ regressor is used only for the preliminary significance test of δ_{MB} presence, and subsequent characterisation of the δ_{MB} time course uses this regressor alone rather than embedded in a difference term.

Principal components analysis. Because preliminary analyses suggested the presence of δ_{MB} and δ_{MF} effects with close temporal and spatial overlap, principal components analysis (PCA) was used as a means of separating these effects. PCA was performed on δ_{MB} and δ_{MF} waveforms using the ERP PCA Toolkit Version 2.63 (Dien 2010a) using similar procedures to those used by Foti, et al. (2011) and Sambrook and Goslin (2016), and following published guidelines (Dien 2010b; Dien, et al. 2005; Dien, et al. 2007). Separate PCAs were performed on δ_{MB} and δ_{MF} waveforms. First, a temporal PCA was performed using each sample point as a variable and each combination of participant and electrode as observations (2880 observations). Factors were retained if they explained more variance than a factor extracted from a null dataset, i.e. they passed a parallel test (Horn, 1965) and were subjected to Promax rotation. All temporal factors that explained more than 1% variance were entered into a spatial PCA in which electrodes were used as variables, and each combination of participant and temporal factor was used as observations. This produced 810 observations for the model-based PCA and 855 for the model-free. Factors that passed a second parallel test were subjected to Infomax rotation. Following the method of Dien, et al. (2003), factors were then reconstructed into waveforms using the product of the factor pattern matrix and the standard deviations. These waveforms, shown in Figure 5, could then be interpreted in the same manner as the original waveforms, in Figure 4, from which they were extracted. While PCA extracts factors based on the variance they explain, without regard to sign, factors that reflect bona fide ERP components should show consistent polarity at the group level. To verify this consistency, one sample t-tests (test value = 0, N = 45) were performed for each factor, scored arbitrarily at its peak amplitude, following Foti et al. (2011).

Rejection of participants. Since we were interested in the neural correlates of reinforcement learning, participants were excluded if they did not learn from outcomes. On a participant-

wise basis we tested for the effect of reward (delivered/omitted) and the interaction of this term with transition type (likely/unlikely) on behaviour on the subsequent trial (stay/switch) following Gillan, et al. (2015). Under this analysis, model-free learning should result in a main effect of reward on stay/switch behaviour and model-based learning should result in an effect of the reward x transition interaction. Four participants showed no significant effect of either term on stay/switch behaviour and were thus excluded from analysis. Participants were also excluded if they were unable to identify which fractals were better when probed at the end of each block. This check was included because of the relatively slow drift of the state 2 – outcome contingencies, which meant that participants who selected the more advantageous state 1 fractal at the outset and then showed strong perseveration could perform at above chance levels without engaging in learning. For each of a participant's blocks, the observed profitability of the state 1 fractals were correlated across blocks with the participant's stated preference of state 1 fractal (left/right) in the end of block probe. An analogous correlation was performed for state 2. Participants were rejected (N = 12) unless they were able to achieve a significant positive correlation for at least one of the two states. It should be noted that rejecting participants on the basis of either their model-based or model-free learning cannot introduce circularity into the study since it is not our intention to show that either kind of learning is typical. Instead, our research question is: in those cases where model based learning occurs, as indexed by behaviour, does it involve computation of a model-based RPEs or not?

Results

Superiority of the hybrid model.

It is generally accepted that humans are capable of model-based reinforcement learning, and the two stage task has previously been successful in demonstrating this. The present study is thus predicated on the assumption that at least some model-based learning will occur and addresses the novel question of whether the EEG shows evidence of a model-based RPE. The success of this hangs on the accuracy with which model-based action values are estimated and these estimates depend on the computational model used. Daw et al. (2011) showed that a hybrid model provided a superior behavioural fit compared to a pure model-based model. To establish this in the present study, each model's free parameters were estimated individually for each participant using maximum likelihood. Based on a total of 32,335 trials, the aggregate raw log likelihood score for the hybrid model was found to be superior to the model-based model (-14,884 vs. -18,870). After taking into account the extra free parameter of omega by converting each log likelihood to its Bayesian Information Criterion, the hybrid model remained superior both in terms of its aggregate level Bayes Factor (BF = 7,676) and the proportion of participants for which the Bayes Factor was superior (1.00). The hybrid model was also superior to a pure model-free model in terms of aggregate raw log likelihood (-15,809), aggregate level Bayes Factor (BF =1,554) and the proportion of participants for which the Bayes Factor was superior (.52). Finally, the Bayesian exceedance probability (Stephan, et al. 2009), or probability that each model is the most common among the three over the population, favoured the hybrid model (hybrid .59, model-based .00, model-free .41).

As a simpler, if cruder, check for model-based learning, we used a mixed effects logistic regression (with random slopes and intercepts fitted to participants) to examine whether stay/switch behaviour on a given trial was predicted both by the previous trial's reward (model-free learning) and its reward x transition interaction (model-based learning). Significance for each fixed effect under consideration was established by comparing the full

model against a null with the fixed effect omitted. This revealed significant effects of both reward ($\chi^2(1) = 99.96$, $p < .001$) and the reward x transition interaction ($\chi^2(1) = 15.16$, $p < .001$) further supporting a hybrid model.

Fitting of behavioural parameters.

Logistic regression, of the model's predicted response probabilities onto the actual response on each trial, produced a significant fit for each participant. We employed Firth's (1993) bias-reduced method, via the *logistf* package (Heinze and Ploner 2016) of R (R Core Team, 2017). Medians for the parameter estimates were: omega 0.39; alpha 0.60; lambda; 0.85; beta 5.77. Omega, a key parameter for the present study, showed a wide range of values (Supplementary Figure 1). As expected, it correlated well with participants' reward x transition coefficient described earlier ($r = .73$, $N = 45$, $p < .001$), and was also correlated with the accuracy with which participants identified the better state 2 ($r = .48$, $N = 45$, $p < .001$). There was no evidence that omega reduced over time as a result of the development of habitual responding, as confirmed by fitting separate omega values for the first and second halves of the experiment and comparing the resulting values (mean ω reduction = .06, paired samples t-test: $t(44) = 1.13$, $p = .25$)

EEGs

Because discriminating model-based and model-free RPEs is problematic owing to their high correlation, and because there was no existing basis for predicting the time or scalp distribution of model-based RPEs, we used a flexible approach to analysis incorporating confirmatory checks, PCA and behavioural covariates. Correction for multiple comparison was implemented in all cases using the Maris and Oostenveld (2007) technique

Outcome locked waveform As a first step, and to establish commonality between the EEG response to the two-step task and other reinforcement learning tasks commonly used in the field, we examined the outcome locked waveform, where only δ_{MF} activity should be observable and should be expected to occur in the interval associated with the feedback related negativity (FRN), from 240 – 340 ms (Sambrook and Goslin 2015). Figure 2 presents the results of the regression V (*voltage)* $= \beta_0 + \beta_1\delta_{MF,2} + \varepsilon$ at FCz, confirming strong sensitivity to δ_{MF} in this interval. Cluster randomisation revealed a single significant cluster running from 200 – 440 ms, initially frontocentral, tending to parietal areas later (Monte Carlo $p < .0001$).

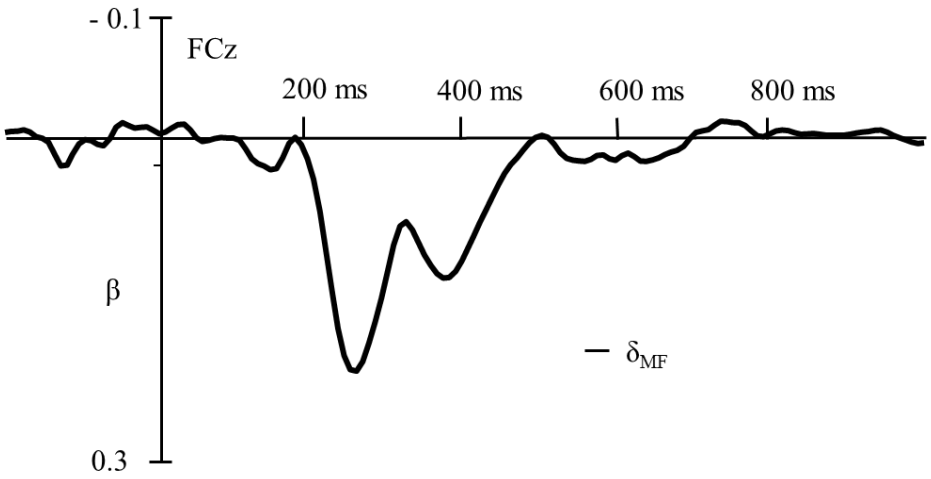

Figure 2. Sensitivity of the outcome locked waveform to model-free RPE, taken from the regression $V = \beta_0 + \beta_1\delta_{MF,2} + \varepsilon$

SPE and Q We next examined effects at state 2. We first assessed the evidence that the SPE and the updated value of state 2, $Q_{MB,2}$ were represented in the EEG, as these are potential

confounds for δ_{MB}. The regression $V = \beta_0 + \beta_1\delta_{MB,1} + \beta_2 SPE + \beta_3 Q_{MB,2} + \varepsilon$ was performed. A significant cluster of SPE related activity was found extending over parieto-occipital sites running from 440 ms to the edge of the measurement window at 1000 ms (Monte Carlo p<.0001, see Supplementary Figure 2). A significant cluster of Q_{MB} related activity was found in frontocentral sites running from 350 – 620 ms (Monte Carlo p = .0013, see Supplementary Figure 3). Additionally, since this regressor is of general interest, we examined Q_{MB} effects in a later window locked to the confirmatory key press and covering the interval -900 – 900 ms, i.e. just before the final outcome. Three clusters of activity were found, one corresponding to that described above for the state 2 window, another parieto-occipital cluster running from -160 – 260 ms (Monte Carlo p < .0001) and a frontocentral cluster from 180 – 780 ms occupying much of the period between the confirmatory response and outcome (Monte Carlo p < .0001). The two frontocentral clusters showed correlated amplitudes across participants at FC4, the site where they were maximal (r = .38, N = 45 p = .01)

Evidence for model-based and model-free RPEs Prior to a regression analysis we performed a preliminary check for RPE encoding by simply comparing average waveforms for positive vs. negative RPEs. This is the standard means of measuring the FRN in one stage tasks, however in the present case RPE sign at state 2 is confounded with the SPE inasmuch as unexpected transitions tend to lead to an undesired state 2. To control for this we performed a 2 x 2 ANOVA using the factors of RPE sign (positive, negative) and SPE (expected, unexpected). RPEs were derived from hybrid Q values to give the best single indicator of RPE activity. The dependent variable was the participant average voltage at FCz in the interval 240 – 340 ms. This revealed a significant main effect of RPE sign ($F_{1,44} = 5.48$, p = .024, $\eta^2 = .11$), with no other effects significant. Figure 3 shows the waveforms. While

present, RPE activity was weak compared to the outcome locked waveform. This is expected because Q values at state 2 are much less extreme than the outcome values of 0 and 1, so tending to produce modest RPEs.

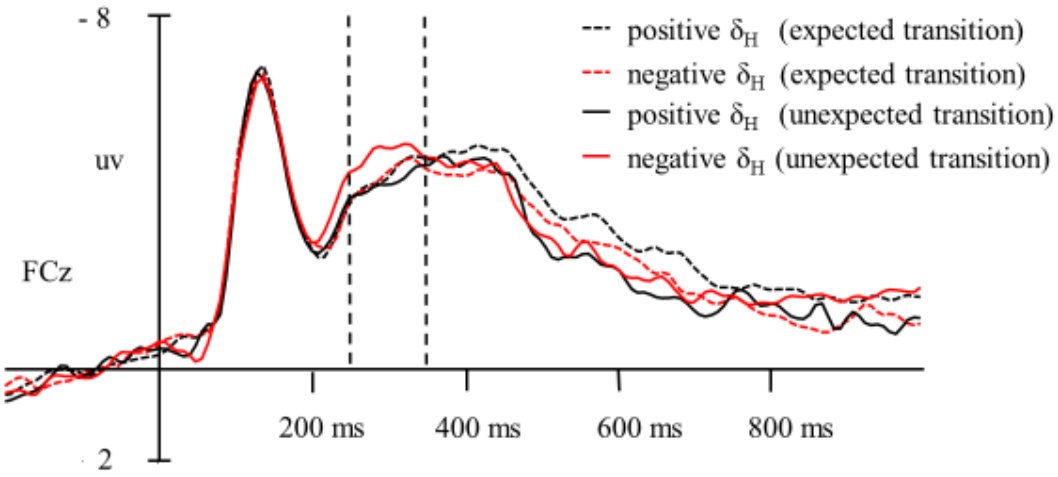

Figure 3. Average waveforms for positive and negative RPEs at state 2

We then progressed to the main purpose of the study, to establish evidence for the existence of a δ_{MB} effect after accounting for the highly correlated δ_{MF} regressor. Following Daw et al., (2011) we performed the regression $V = \beta_0 + \beta_1\delta_{MF,1} + \beta_2(\delta_{MB,1} - \delta_{MF,1}) + \beta_3 SPE + \beta_4 Q_{MF,2} + \varepsilon$. The $\delta_{MB} - \delta_{MF}$ regressor produced a significant cluster of activity from 310 – 660 ms at parietal sites (Monte Carlo p = .002) indicating residual effects of δ_{MB} once those from δ_{MF} had been accounted for. As an alternative test for δ_{MB} effects, we compared the adjusted R^2 values of two models: $V = \beta_0 + \beta_1\delta_{MF,1} + \beta_2 SPE + \beta_3 Q_{MF,2} + \varepsilon$ vs. $V = \beta_0 + \beta_1\delta_{MF,1} + \beta_2\delta_{MB,1} + \beta_3 SPE + \beta_4 Q_{MF,2} + \varepsilon$. A significant improvement of adjusted R^2 in the second model would indicate the presence of a δ_{MB} effect; in comparison, if no such effect were present, adjusted R^2 should be worse as a result of downward adjustment due to a non-explanatory second regressor. Note that since individual betas were not inspected, the collinearity arising

from the presence of both δ_{MF} and δ_{MB} is of no concern here. In line with the previous analysis, this comparison revealed a significant improvement in adjusted R^2 values for the model containing δ_{MB}, running from 360 – 420 ms at centroparietal sites.

Temporospatial character of model-based and model-free RPEs As noted earlier, while the residual $\delta_{MB} - \delta_{MF}$ is useful for testing for the presence of a δ_{MB} effect in the context of a correlated δ_{MF} regressor, the temporospatial character of the δ_{MB} effect is not well expressed in this form. Thus having established evidence that δ_{MB} effects were present, we reverted to the simpler model $V = \beta_0 + \beta_1\delta_{MB,1} + \beta_2 SPE + \beta_3 Q_{MB,2} + \varepsilon$ to assess the temporospatial character of δ_{MB}. This revealed a single significant cluster of δ_{MB} activity running from 210 – 350 ms over frontocentral sites (Monte Carlo p =.0003). The complementary model $V = \beta_0 + \beta_1\delta_{MF,1} + \beta_2 SPE + \beta_3 Q_{MF,2} + \varepsilon$ revealed a single cluster of δ_{MF} activity running from 200 – 420 ms, initially maximal frontocentrally, progressing to parietal areas later (Monte Carlo p <.0001). As can be seen in Figure 4, in the early interval from 200 – 300 ms, voltage is better predicted by δ_{MB} than δ_{MF} activity.

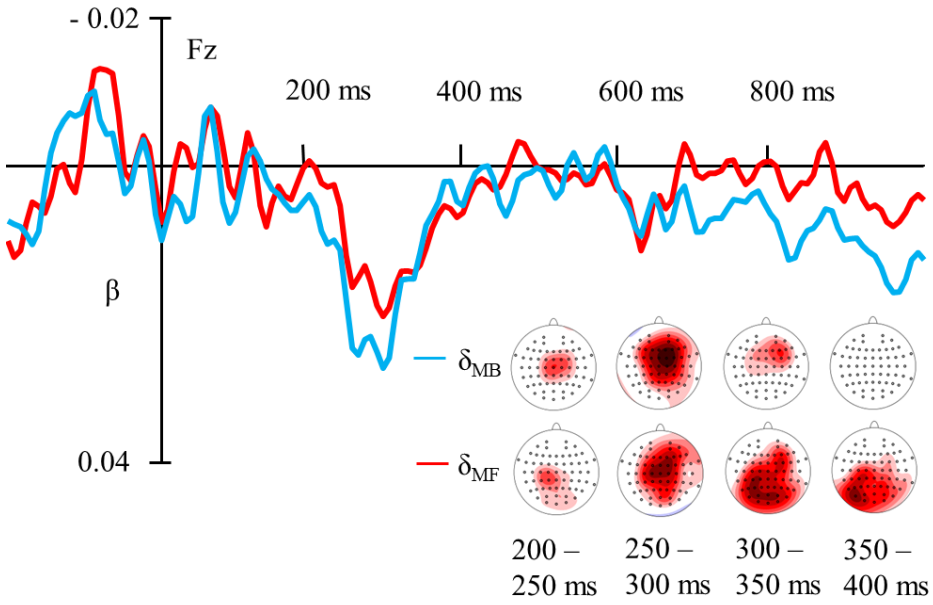

Figure 4. Sensitivity of the state 2 waveform to model-based and model-free RPEs taken from the regressions $V = \beta_0 + \beta_1\delta_{MB,1} + \beta_2 SPE + \beta_3 Q_{MB,2} + \varepsilon$ and $V = \beta_0 + \beta_1\delta_{MF,1} + \beta_2 SPE + \beta_3 Q_{MF,2} + \varepsilon$ respectively

In order to better separate δ_{MB} and δ_{MF} effects, these coefficients were subjected to PCA. For the δ_{MB} coefficient, eight factors accounted for greater than 1% of the overall variance, but of these only one factor (TS4/SF1) was sufficiently consistent in its polarity across participants to be significant under a t test ($t(44) = 3.13$, $p = .003$). This factor extended over frontocentral and parietal sites, peaking at 216 ms. For the δ_{MF} coefficient, sixteen factors accounted for greater than 1% of the overall variance, but again only one (TS4/SF1) achieved significance under a t test ($t(44) = 3.02$, $p = .004$). This factor peaked at 296 ms and was more centroparietal. The factors are shown in Figure 5. While they broadly correspond to the most prominent areas of effect of the original coefficients as shown in Figure 4, there is a marked degree of temporal separation after PCA, confirming these as

being distinct components and providing further evidence for an early frontocentral model-based RPE.

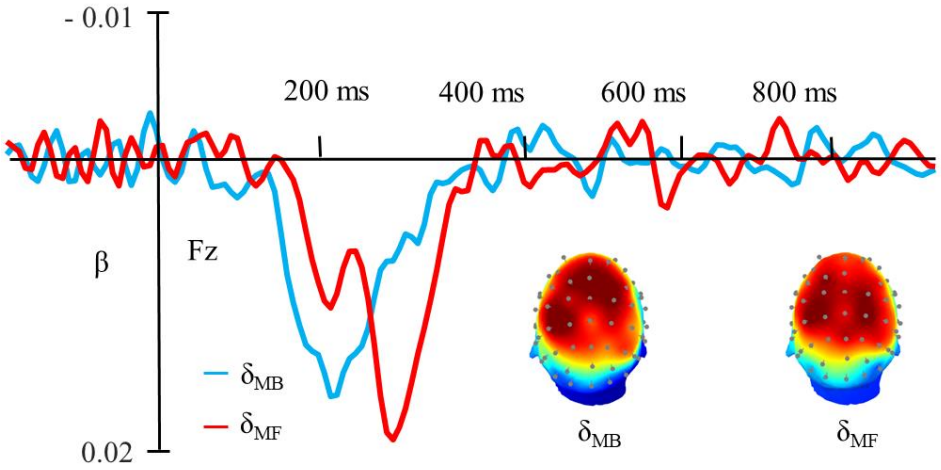

Figure 5. Temporospatial factors derived from PCA on δ_{MB} and δ_{MF} effects shown in Figure 4

An alternative means of separating δ_{MB} effects from overlapping δ_{MF} effects is to correlate participants' δ_{MB} effect with their omega estimate (Daw, et al. 2011). Insofar as δ_{MB} effects should be stronger when a participant is engaged in model-based learning, this correlation should serve to accentuate the δ_{MB} signal wherever it lies, and pare away apparent δ_{MB} effects that in reality reflect δ_{MF} activity. We therefore preformed the regression $\delta_{MB,1} = \beta_0 + \beta_1\omega + \varepsilon$. This was necessarily performed across participants rather than for participants individually. For Monte Carlo testing, null datasets were achieved by random rearrangement of ω values over participants prior to regression. The previously identified early δ_{MB} effect at 210 – 350 ms showed a positive correlation with omega but failed to retain significance after correction for multiple comparisons. Instead, a strongly significant cluster of activity was found running from 450 – 700 ms at centroparietal sites (Monte Carlo p = .0003). The complementary model, $\delta_{MF,1} = \beta_0 + \beta_1\omega + \varepsilon$, revealed no significant clusters. These effects are

shown in Figure 6 at CP4 where the effect of model-based RPE as a function of omega was maximal.

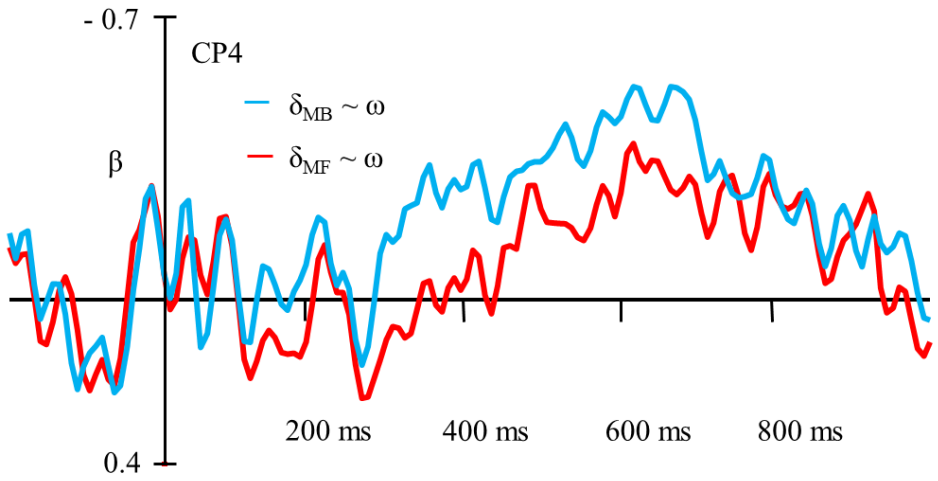

Figure 6. Sensitivity of δ_{MB} and δ_{MF} effects shown in Figure 4 to the behavioural estimate of model-based learning, ω

Evidence for computationally separate model-based and model-free RPEs The evidence for δ_{MB} and δ_{MF} effects has so far been taken to imply separate, computationally encapsulated systems. However, rather than reflecting specific learning signals of this sort, activity in the feedback locked waveform has also been suggested to reflect rather more general expectancy violation, such as that related to affect (Gehring and Willoughby 2002; Luu, et al. 2003) or conflict (Folstein and Van Petten 2008). Insofar as both model-based and model-free learning will have contributed to this expectancy, an EEG response scaled to its violation will show an incidental correlation with the δ_{MB} and δ_{MF} terms of our model even if these actual terms are not computed as such by the brain. We tested between these two interpretations. Generalised reward expectancy was operationalised by the Q_H term, since this is a behaviourally derived estimate of participants' overall expected value for the action taken. The violation of overall

reward expectation can thus be operationalised with the associated RPE, δ_H. We should be careful to note that we do not hypothesise a computational use for δ_H in reinforcement learning, we merely use it here as an index of general reward expectancy violation. We compared the adjusted R^2 values of two models: $V = \beta_0 + \beta_1\delta_{MB,1} + \beta_2\delta_{MF,1} + \beta_3 SPE + \beta_4 Q_{MB,2} + \varepsilon$ and $V = \beta_0 + \beta_1\delta_{H,1} + \beta_2 SPE + \beta_3 Q_{MB,2} + \varepsilon$. If the EEG merely reflects deviation from generalised reward expectation, then the model containing δ_H should produce the strongest effect size. In contrast, if separate RPEs are generated by model-based and model-free learning modules, then a model with these entered as individual regressors will be superior. Comparison of adjusted R^2 values revealed the δ_H model to be generally inferior to the model containing separate δ_{MB} and δ_{MF} terms, with the effect most pronounced at a significant cluster of centroparietal sites from 300 – 440 ms (Monte Carlo p = .0002).

Positive vs. negative RPEs. RPEs in our models are signed terms, and the regressions incorporate this property arithmetically without according it any particular relevance, thus deriving betas based on a linear relationship of voltage to RPE across this variable's full bivalent range. This may be an inappropriate assumption inasmuch as some studies have claimed that instrumental learning tasks produce RPE activity in the EEG that is driven selectively by negative (Bellebaum and Daum 2008; Cohen, et al. 2007) or positive (Foti, et al. 2011; Liu, et al. 2014; Sambrook and Goslin 2016) RPEs. The question can be investigated by regressing voltage against positive and negative RPEs separately, establishing regions of sensitivity, and comparing the sign of the respective betas in these regions. A bivalent encoding of RPEs is implied when the models produce same-signed betas, a univalent encoding (such as claimed by the authors above) when betas are significant for one valence only, and an unsigned prediction error encoding (sometimes described as salience, Talmi et al., 2013) when betas are oppositely signed. If valences are not separately modelled

(as is the case here) the consequence will be that components encoding unsigned prediction error will be removed from the EEG through cancelling, univalent components will remain, though presenting a shallower beta than when modelled separately by valence (and generally showing weaker significance, the increased sample size notwithstanding) and bivalent components will be maintained with increased significance. Our objective in the present study was to establish if *any* kind of δ_{MB} activity occurred and, without any basis for predicting valence-specific sensitivity, we assumed the simplest, bivalent, model. This was partly to maintain sample size in the face of effects that were necessarily small when analysis was performed on state 2 feedback and partly for robustness of interpretation given that, as indicated above, such an analysis always retains all bona fide RPE activity and removes all unsigned prediction error activity. However, having demonstrated the presence of model-based RPE activity, we re-ran models for positive and negatively valenced RPEs separately, to provide additional insight into the effects shown in Figures 3 and 5. For both δ_{MF} or δ_{MB} waveforms this decomposition revealed no significant difference between positive and negative betas in frontocentral sites from 200 - 420 ms, nor for omega correlations at centro-parietal sites from 450 – 700 ms. Betas were generally same-signed in these intervals, however the reduced power makes it difficult to discriminate between bivalent and univalent codings. Waveforms are provided in Supplementary Figures 4 and 5.

Discussion

The present study provides evidence for model-based RPEs in the EEG. This was based both on the ability of a residual term, $\delta_{MB} - \delta_{MF}$, to account for variance not explained by model-

free RPEs, and also by a comparison of the overall variance explained by a model containing model-based and model-free RPEs compared to one containing only model-free RPEs.

An early model-based RPE, isolated by PCA at 216 ms at frontocentral sites, was distinguishable from a model-free RPE signal peaking some 80 ms later. This model-based RPE was uncorrelated with omega, the degree to which behaviour was model-based. Such dissociations of learning signals and behaviour are not uncommon. Bayer and Glimcher (2005) demonstrated that model-free RPEs continued to be faithfully computed even when a monkey was pursuing a quite different model-based behavioural policy and, using EEG, Chase, et al. (2011) showed that in a reversal learning task, FRN amplitudes reflected model-free RPEs to old value estimates even when participants' behaviour indicated that they believed values had been reversed. Conversely, Walsh and Anderson (2011) showed that the FRN requires exposure to reinforcement in order to develop (a hallmark of model-free learning) despite behaviour reflecting the immediate adoption of a model-based policy. In all these cases, model-free neural signals are computed despite behaviour being model-based. Here we show the reverse: that model-based RPEs are computed even if the participant's behaviour is model-free. It may initially seem paradoxical that an agent should maintain a model of the environment on which they do not act. However, this is entirely consistent with the contemporary perspective of model-free and model-based learning systems running in parallel and competing for access to behaviour (Daw and O' Doherty 2013) and the apparent paradox simply arises from the fact that model-free forms of prediction and control are largely hidden from explicit subjective view (Huys, et al. in press).

Both model-based and model-free RPEs were present in the interval from 240 – 340 ms, the time at which the FRN occurs. The FRN has been claimed to code a model-free RPE (Chase, et al. 2011; Walsh and Anderson 2011) and it is likely that this is at least in part true. Some studies however, have shown activity associated with the FRN which appears to reflect

wider knowledge. Reiter, et al. (2016) showed that the FRN appears to incorporate inferences about how values of an unchosen key have changed, and Collins and Frank (2016) found evidence that latent rule structure influenced the FRN. Given the wide interval in which this component is measured, it is likely that the FRN reflects a composite of several neural generators (Foti, et al. 2014), with its apparent character (e.g. model-free vs. model-based) dependent on when and how it is operationalised. This stresses the value of a fully data-driven means of extracting effects such as that used here, rather than the use of a pre-defined canonical interval selected from a variety of such intervals available in the literature.

The common scalp topography of the model-free and model-based effects found here is consistent with Daw et al.'s (2011) claim of a common substrate for the two forms of learning. Source localisation was not attempted in the present study because the small effect sizes associated with state 2 RPEs result in an unstable solution. However we have previously localised frontocentral scalp distributions such as those shown by model-based and model-free RPE here, to the striatum (Sambrook and Goslin 2016), the same common substrate found by Daw et al. While the striatum is regarded as key site of model-free learning, other studies have also located model-based effects there (FitzGerald, et al. 2010).

A late, negative-going centroparietal effect of model-based RPE was also found. In contrast to the early model-based effect this was heavily correlated with omega: indeed it could only be detected once this covariate was included. As such, this effect appears to reflect feedback processing in the context of expressed behaviour, rather than the automatic tracking and revision of model-based expected value implied by the early component. Such behaviour-linked effects have been demonstrated before, for example Chase, et al. (2011) showed that negative RPEs which preceded rule-based behavioural reversal were associated with greater P3 amplitude than equivalent sized ones which did not. However it is also possible that the late effect shown here reflects more generalised processing of valence rather

than model-based RPE as such, and is larger in participants pursuing a model-based strategy simply because they were more invested in the task. The effect's late latency and sustained character are consistent with more deliberative processes. Additionally, the outcome locked waveform, where only model-free RPEs are possible, revealed a very similar negative going centroparietal effect (Monte Carlo $p = .15$).

The demonstration of model-based RPEs does not resolve the question of what function they serve. However, the finding of a model-based RPE at an early latency, computed regardless of a participant's overt behaviour, suggests that model-based RPEs may serve a fundamental computational role in reinforcement learning. The early latency, possibly up to 80 ms earlier than model-free RPE computation, may seem surprising given that model-based processes are regarded as computationally costly, but this cost only applies to the *generation* of valuation, not the RPE generated against it, which is equally trivial for both model-free and model-based learning. It is this trivial cost that should lead us to the view that if model-based RPEs can serve a function then the brain will compute them. The possible co-location of model-based and model-free RPEs in the striatum is certainly consistent with the view that model-based RPEs are used in model-free updates, as has been proposed by Daw, et al. (2011). Where the model is good this will generally lead to superior model-free performance. Performance will still be worse than a full model-based system, but will still be beneficial if model-based valuation is interrupted, for example due to cognitive load. Thus the development of computational models describing the interaction of model-free and model-based learning are needed to assess their importance in determining behaviour in the normal case. The terms from such models can also be combined with neural data to shed light on the computational substrate of this interaction. This study, and that of Daw et al.'s, have brought converging evidence to bear using fMRI and EEG, but there is also evidence that

single cells can show behaviour that incorporates information beyond simple model-free RPEs (Bromberg-Martin, et al. 2010; Nakahara, et al. 2004).

One impetus for using EEG to identify model-based RPEs was that its temporal precision allows RPEs to be disentangled from action values, both before and after the RPE update. How action values are represented in the EEG is of interest in its own right however. Here, using single trial parametric testing on computationally modelled reward prediction errors, whilst controlling for confounding components, we were also able to reveal action values in the EEG. One cluster of activity occurred frontocentrally directly after the model-free prediction error. Studies by Tzovara, et al. (2015), Fischer and Ullsperger (2013) and Hunt, et al. (2012) (using MEG), employing tasks in which variably valuable choices were presented at the start of a trial have also successfully demonstrated neural signatures of Q values. Also, in an experiment employing the same two step task as that used here, Eppinger, et al. (2017) showed state action Q values at state 2 at a comparable latency (400 ms) to ourselves, though more parietally. The speed of onset following state 2 feedback suggests a relatively automatic use of RPEs to update model-free action values.

In the course of removing its confounding effects, we also showed that unexpected transitions resulted in a large sustained positive parietal potential from 400 – 800 ms. Activity in this time and location is sometimes referred to as the Slow Wave or Late Positive Complex. Surprising, or otherwise salient outcomes, have previously been associated with this component. Spencer, et al. (2001) showed such a response to oddball stimuli and Sambrook and Goslin (2014, 2016) observed such a response scaled to the absolute size of RPEs regardless of whether these were manipulated by outcome likelihood or magnitude. Eppinger et al 2017, using the same two step task also found a positive going parietal component associated with SPEs, maximal at 800 ms. Precisely what aspects of salience this component reflects is yet to be resolved however. In the context of reinforcement learning,

the term salience has been used to refer to a number of different properties including infrequency (Schultz 2013), z-score (Schultz 2009), unsigned prediction error (Sambrook and Goslin 2014; Sambrook and Goslin 2016; Talmi, et al. 2013), mere presence or absence of a stimulus (Esber and Haselgrove 2011) or, in the present case, SPE, and distinguishing between these will require careful experimental design. However, Glascher, et al. (2010), using the current two-step task but with fMRI, showed activation that was better attributed to a SPE than an unsigned prediction error, with this effect present in areas associated with the P3, a component that is widely regarded as reflecting model updating (Donchin and Coles 1988). Cavanagh (2015), using a three armed bandit task, and a computational model incorporating SPE, Q and δ_{MF} terms, much like ours, also showed the P3 reflected behavioural adjustment associated with an SPE, with this dissociated from δ_{MF} activity shown in the time course of the FRN. The effect reported in the present study overlaps with the time-course of the P3 and it is very possible that a SPE is being recorded in both Glascher et al.'s and our study. However this question will be best addressed by explicitly modelling SPEs in a more dynamic learning task where participants must monitor contingency changes. This contrasts with the present case where participants were instructed to regard contingencies as fixed.

To conclude, this study identifies neural correlates of the computational building blocks of reinforcement learning. It shows the usefulness of multiple regression of single trial EEG data for separating out computational terms that are correlated in typical reinforcement learning experiments. The demonstration of a model-based RPE using EEG, in concert with previous evidence drawn from fMRI, suggests that the current paradigm of fully dissociated model-free and model-based systems may not be realistic. This might entail a modification of current concepts of model-based valuation to incorporate a model-based RPE, or may instead require a revision of model-free learning such that it receives input from the model-based

system via an RPE. These new computational models can then be submitted for empirical verification. In the wider context, elucidating the mechanisms of model-free and model-based learning is central to understanding how harmful habits are formed and maintained. A strong model-based learning system can serve as a protective factor against psychiatric conditions of habit such as drug addiction, eating and anxiety disorders (Hasler 2012). The field of computational psychiatry (Maia and Frank 2011; Montague, et al. 2012; Redish 2004) seeks to use computational terms from models of normative decision making as biomarkers of mental disorders in the case where those terms have outlying values. Thus Paulus, et al. (2004) showed that, when expressed as computational terms, both behaviour and fMRI activations elicited from a simple two-choice gambling task could be used to predict relapse in a cohort of treatment-seeking methamphetamine addicts. Using the current two-step task but with a devaluation manipulation at the task's conclusion, Gillan, et al. (2015) showed that the behavioural term omega predicted individual-wise devaluation sensitivity. Model-based learning thus appears to be key to resisting habit formation, and measuring its strength may predict important real world adaptive behaviour such as sensitivity to reward devaluation or contingency changes. The availability of neural in addition to behavioural predictors can only help in this regard, not least since the two may be dissociated as in the case of the early model-based RPE shown here. It might also be argued that the predictive value of neural biomarkers for real world behaviour will be more reliable than that of behavioural assays insofar as performance on any particular experimental task may generate heavily task led behaviour peculiar to that circumstance. In contrast, the strength of the response of the neural apparatus activated may more robustly predict that apparatus' involvement in the wider context of the real world.

References

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47*(1), 129-141. doi:DOI 10.1016/j.neuron.2005.05.020

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience, 27*(7), 1823-1835. doi:DOI 10.1111/j.1460-9568.2008.06138.x

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values. *Journal of Neurophysiology, 104*(2), 1068-1076. doi:10.1152/jn.00158.2010

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Cavanagh, J. F. (2015). Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. *Neuroimage, 110*, 205-216. doi:10.1016/j.neuroimage.2015.02.007

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related Negativity Codes Prediction Error but Not Behavioral Adjustment during Probabilistic Reversal Learning. *Journal of Cognitive Neuroscience, 23*(4), 936-946. doi:DOI 10.1162/jocn.2010.21456

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage, 35*(2), 968-978. doi:DOI 10.1016/j.neuroimage.2006.11.056

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition, 152*, 160-169. doi:10.1016/j.cognition.2016.04.002

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., & O' Doherty, J. P. (2013). Multiple systems for value learning. In P. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (Second ed., pp. 393-410). San Diego: Elsevier.

de Wit, S., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: a case for animal-human translational models. *Psychological Research-Psychologische Forschung, 73*(4), 463-476. doi:10.1007/s00426-009-0230-6

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., & Ridderinkhof, K. R. (2012). Corticostriatal Connectivity Underlies Individual Differences in the Balance between Habitual and Goal-Directed Action Control. *Journal of Neuroscience, 32*(35), 12066-12075. doi:10.1523/Jneurosci.1088-12.2012

Dickinson, A. (1985). Actions and Habits - the Development of Behavioral Autonomy. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 308*(1135), 67-78. doi:DOI 10.1098/rstb.1985.0010

Dien, J. (2010a). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods, 187*(1), 138-145. doi:DOI 10.1016/j.jneumeth.2009.12.009

Dien, J. (2010b). Evaluating two-step PCA of ERP data with Geomin, Infomax, Oblimin, Promax, and Varimax rotations. *Psychophysiology, 47*(1), 170-183. doi:DOI 10.1111/j.1469-8986.2009.00885.x

Dien, J., Beal, D. J., & Berg, P. (2005). Optimizing principal components analysis of event-related potentials: Matrix type, factor loading weighting, extraction, and rotations. *Clinical Neurophysiology, 116*(8), 1808-1825. doi:DOI 10.1016/j.clinph.2004.11.025

Dien, J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax vs. infomax rotations. *Human Brain Mapping, 28*(8), 742-763. doi:Doi 10.1002/Hbm.20304

Dien, J., Spencer, K. M., & Donchin, E. (2003). Localization of the event-related potential novelty response as defined by principal components analysis. *Cognitive Brain Research, 17*(3), 637-650. doi:Doi 10.1016/S0926-6410(03)00188-5

Donchin, E., & Coles, M. G. H. (1988). Is the P300 Component a Manifestation of Context Updating. *Behavioral and Brain Sciences, 11*(3), 357-374.

Eppinger, B., Walter, M., Heekeren, H. R., & Li, S. C. (2013). Of goals and habits: age-related and individual differences in goal-directed decision-making. *Front Neurosci, 7*. doi:ARTN 253 10.3389/fnins.2013.00253

Eppinger, B., Walter, M., & Li, S. C. (2017). Electrophysiological correlates reflect the integration of model-based and model-free decision information. *Cognitive Affective & Behavioral Neuroscience, 17*(2), 406-421. doi:10.3758/s13415-016-0487-3

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B-Biological Sciences, 278*(1718), 2553-2561. doi:DOI 10.1098/rspb.2011.0836

Firth, D. (1993). Bias Reduction of Maximum-Likelihood-Estimates. *Biometrika, 80*(1), 27-38. doi:DOI 10.1093/biomet/80.1.27

Fischer, A. G., & Ullsperger, M. (2013). Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron, 79*(6), 1243-1255. doi:10.1016/j.neuron.2013.07.006

FitzGerald, T. H. B., Seymour, B., Bach, D. R., & Dolan, R. J. (2010). Differentiable Neural Substrates for Learned and Described Value and Risk. *Current Biology, 20*(20), 1823-1829. doi:10.1016/j.cub.2010.08.048

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology, 45*(1), 152-170. doi:10.1111/j.1469-8986.2007.00602.x

Foti, D., Weinberg, A., Bernat, E. M., & Proudfit, G. H. (2014). Anterior cingulate activity to monetary loss and basal ganglia activity to monetary gain uniquely contribute to the feedback negativity. *Clinical Neurophysiology*.

Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-Related Potential Activity in the Basal Ganglia Differentiates Rewards from Nonrewards: Temporospatial Principal Components Analysis and Source Localization of the Feedback Negativity. *Human Brain Mapping, 32*(12), 2207-2216. doi:Doi 10.1002/Hbm.21182

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295*(5563), 2279-2282.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive Affective & Behavioral Neuroscience, 15*(3), 523-536. doi:10.3758/s13415-015-0347-6

Glascher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron, 66*(4), 585-595. doi:10.1016/j.neuron.2010.04.016

Glimcher, P. W. (2009). Choice: Towards a standard back-pocket model. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making*

and the brain (First edition ed., pp. 538-521). San Diego, CA: Elsevier Academic Press.

Hald, A. (2003). *History of Probability and Statistics and Their Applications before 1750*. Hoboken, New Jersey.: John Wiley & Sons.

Hasler, G. (2012). Can the neuroeconomics revolution revolutionize psychiatry? *Neuroscience and Biobehavioral Reviews, 36*(1), 64-78. doi:10.1016/j.neubiorev.2011.04.011

Heinze, G., & Ploner, M. (2016). logistf: Firth's Bias-Reduced Logistic Regression. R package version 1.22.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679-709. doi:Doi 10.1037//0033-295x.109.4.679

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology, 45*(5), 688-697. doi:DOI 10.1111/j.1469-8986.2008.00668.x

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience, 15*(3), 470-U169. doi:10.1038/nn.3017

Huys, Q. J., Beck, A., Dayan, P., & Heinz, A. (in press). Neurobiology and computational structure of decision-making in addiction. In M. e. al (Ed.), *Phenomenological Neuropsychiatry: Bridging the Clinic and Clinical Neuroscience*.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex, 13*(4), 400-408. doi:DOI 10.1093/cercor/13.4.400

Liu, W. H., Wang, L. Z., Shang, H. R., Shen, Y., Li, Z., Cheung, E. F. C., & Chan, R. C. K. (2014). The influence of anhedonia on feedback negativity in major depressive disorder. *Neuropsychologia, 53*, 213-220. doi:10.1016/j.neuropsychologia.2013.11.023

Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science, 14*(1), 47-53. doi:Doi 10.1111/1467-9280.01417

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience, 14*(2), 154-162. doi:10.1038/nn.2723

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190. doi:DOI 10.1016/j.jneumeth.2007.03.024

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences, 16*(1), 72-80. doi:10.1016/j.tics.2011.11.018

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41*(2), 269-280. doi:Doi 10.1016/S0896-6273(03)00869-9

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. J., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452-454. doi:DOI 10.1126/science.1094285

Paulus, M. P., Schuckit, M. A., & Tapert, S. F. (2004). Neural activation patterns of methamphetamine dependent subjects during decision-making predict. *Neuropsychopharmacology, 29*, S27-S27.

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical Splines for Scalp Potential and Current-Density Mapping. *Electroencephalography and Clinical Neurophysiology, 72*(2), 184-187. doi:Doi 10.1016/0013-4694(89)90180-6

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science, 306*(5703), 1944-1947. doi:10.1126/science.1102384

Reiter, A. M. F., Koch, S. P., Schroger, E., Hinrichs, H., Heinze, H. J., Deserno, L., & Schlagenhauf, F. (2016). The Feedback-related Negativity Codes Components of Abstract Inference during Reward-based Decision-making. *Journal of Cognitive Neuroscience, 28*(8), 1127-1138. doi:10.1162/jocn_a_00957

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory, 2*, 64-99.

Sambrook, T. D., & Goslin, J. (2014). Mediofrontal event-related potentials in response to positive, negative and unsigned prediction errors. *Neuropsychologia, 61*, 1-10. doi:DOI 10.1016/j.neuropsychologia.2014.06.004

Sambrook, T. D., & Goslin, J. (2015). A Neural Reward Prediction Error Revealed by a Meta-Analysis of ERPs Using Great Grand Averages. *Psychological Bulletin, 141*(1), 213-235. doi:Doi 10.1037/Bul0000006

Sambrook, T. D., & Goslin, J. (2016). Principal components analysis of reward prediction errors in a reinforcement learning task. *Neuroimage, 124*, 276-286. doi:10.1016/j.neuroimage.2015.07.032

Schultz, W. (2009). Midbrain dopamine neurons: a retina of the reward system. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (First edition ed., pp. 323-329). San Diego, CA: Elsevier Academic Press.

Schultz, W. (2013). Updating dopamine reward signals. *Current Opinion in Neurobiology, 23*(2), 229-238. doi:10.1016/j.conb.2012.11.012

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin Selectively Modulates Reward Value in Human Decision-Making. *Journal of Neuroscience, 32*(17), 5833-5842. doi:10.1523/Jneurosci.0053-12.2012

Spencer, K. M., Dien, J., & Donchin, E. (2001). Spatiotemporal analysis of the late ERP responses to deviant stimuli. *Psychophysiology, 38*(2), 343-358. doi:Doi 10.1017/S0048577201000324

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage, 46*(4), 1004-1017. doi:10.1016/j.neuroimage.2009.03.025

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*: MIT Press.

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The Feedback-Related Negativity Signals Salience Prediction Errors, Not Reward Prediction Errors. *Journal of Neuroscience, 33*(19), 8264-8269. doi:Doi 10.1523/Jneurosci.5695-12.2013

Tzovara, A., Chavarriaga, R., & De Lucia, M. (2015). Quantifying the time for accurate EEG decoding of single value-based decisions. *Journal of Neuroscience Methods, 250*, 114-125. doi:10.1016/j.jneumeth.2014.09.029

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences of the United States of America, 108*(47), 19048-19053. doi:DOI 10.1073/pnas.1117189108

Wills, A. J., O'Connell, G., Edmunds, C. E. R., & Inkster, A. B. (2017). Progress in Modeling Through Distributed Collaboration: Concepts, Tools and Category-Learning Examples. *Psychology of Learning and Motivation, Vol 66, 66*, 79-115. doi:10.1016/bs.plm.2016.11.007

Wills, A. J., & Pothos, E. M. (2012). On the Adequacy of Current Empirical Evaluations of Formal Models of Categorization. *Psychological Bulletin, 138*(1), 102-125. doi:10.1037/a0025715

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19*(1), 181-189. doi:10.1111/j.1460-9568.2004.03095.x

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience, 22*(2), 513-523. doi:10.1111/j.1460-9568.2005.04218.x