How instructed knowledge shapes adaptive learning

Lauren Y. Atlas[1,2*], Bradley B. Doll[3,4], Jian Li[5,6], Nathaniel D. Daw[7,8], Elizabeth A. Phelps[3,9,10]

[1] National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, MD
[2] National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD
[3] Center for Neural Sciences, New York University, New York, NY
[4] Department of Psychology, Columbia University, New York, NY
[5] Department of Psychology and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China
[6]PKU-IDG/McGovern Institute for Brain Research, Peking University
[7] Princeton Neuroscience Institute, Princeton University, Princeton, NJ
[8] Department of Psychology, Princeton University, Princeton, NJ
[9] Department of Psychology, New York University, New York, NY
[10] Nathan Kline Institute, Orangeburg, NY

* Please address correspondence to:

Lauren Y. Atlas, Ph.D.
National Center for Complementary and Integrative Health/ National Institutes of Health
10 Center Drive, Rm 4-1741
Bethesda, MD 20892
Email: lauren.atlas@nih.gov
Phone: 301-827-0214

Running Title: HOW INSTRUCTIONS SHAPE ADAPTIVE LEARNING

Abstract

We previously published results of a study examining how instructions shape aversive reversal learning (Atlas et al., 2016). Our original paper measured how instructions and learning influence on expected value, and assumed that learning rates remain stable over time. We found that instructions caused immediate reversals in corticostriatal systems, while value signals in the amygdala required reinforcement in order to update. However, our paradigm was originally designed to measure adaptive learning, and to test whether previously observed neural dissociations in the encoding of associability and prediction error i) replicate in the context of repeated reversals and ii) are modulated with instructions. Here, we present reanalyses using a hybrid model, in which associability gates learning. In our Uninstructed Group, we replicate previous dissociations (Li et al., 2011) such that the striatum tracks prediction error while the amygdala tracks associability. Comparisons with an Instructed Group reveal that instructions update expected value and striatal prediction error, while amygdala associability updates based on feedback alone, irrespective of instruction. This work adds to a growing body of literature on adaptive learning and provides new results that suggest that instructions have dissociable effects on prediction error and associability.

Computational models of feedback-driven associative learning reveal that coordinated brain systems update expectations dynamically in response to rewards and punishments in the environment [1]. We and others have recently shown that instructions modulate responses in corticostriatal systems during both appetitive and aversive learning [2–4]. Our previous analyses focused on expected value (EV) and revealed that EV signals in the striatum, orbitofrontal/ventromedial prefrontal cortex (OFC/VMPFC), and salience network update upon instruction, whereas EV signals in the bilateral amygdala require reinforcement in order to update[2]. The use of a static learning rate allowed us to test directly for dissociations using a single parameter (EV). However, learning is often adaptive and the rate of learning can vary depending on features of the environment, particularly when environmental contingencies change [5–7]. In fact, our study was originally designed to measure how adaptive learning is modulated by instructions during aversive reversal learning.

We focused on how instructions interact with error-driven learning in a model (the "hybrid" model of [5]) in which incremental learning from feedback is modulated by associability. Associability describes the process whereby the speed of learning decreases as an environment becomes predictable, and increases when unexpected changes occur [8]. Associability can account for processes that influence learning such as attention, habituation, and changes in context, and is related to other statistically motivated modulations of learning that have been evaluated in human neuroimaging studies [7,9,10]. Several studies [5,6,11] have shown that this adaptive model characterizes human behavior and blood oxygenation level dependent (BOLD) activity during aversive reversal learning better than an error-driven learning model with a fixed learning rate (i.e. Rescorla-Wagner without associability). For example, the hybrid model was fit to skin conductance responses (SCRs), a traditional measure of the conditioned fear response in humans, measured during a fear conditioning study with a single reversal [5,12]. FMRI analyses revealed dissociations in the brain systems that tracked dynamic learning, such that the ventral striatum was correlated with dynamic prediction errors (PEs), whereas the amygdala tracked associability [5].

It is unknown whether adaptive aversive learning is modulated when individuals are informed about contingencies in the environment. Several studies in the appetitive domain suggest that instructions about rewarding outcomes modulate learning-related responses (e.g. PEs) in the striatum and that this modulation might depend on the prefrontal cortex [3,4,13]. However, no studies have tested whether instructions have the same effects on adaptive learning. Instructions might modulate associability in the amygdala, as well as striatal PEs, or amygdala responses might be insensitive to cognitive instruction. We previously used a simpler model that assumed that learning remained stable over time [2] and found that expected value signals (EV) in the striatum, OFC/VMPFC, and salience network updated upon instruction, while EV signals in the amygdala followed the course of feedback-driven learning, based on fits to an Uninstructed Group. This model was applied because it was parsimonious and allowed us to test dissociations using a single computational parameter (EV). Yet reversal learning and learning in volatile environments may be better described by models that take into account environmental uncertainty and include variations in learning over time based on uncertainty in the environment. In the present reanalysis, we tested whether similar dissociations were observed when we used a hybrid model that incorporated associability.

Results

Feedback-driven learning is modulated by instructions. We fit a family of reinforcement learning models to the Uninstructed Group's SCRs to isolate feedback-driven aversive learning in the presence of serial reversals. This isolates learning-related processes that respond to reinforcement history alone. Shocks were incorporated as reinforcements, and thus a positive expected value (EV) corresponds to an expectation for a shock. EV updates in response to prediction errors (PEs), and the speed at which EV updates depends on learning rate (see Methods). We compared a model with a fixed learning rate (i.e. a basic Rescorla-Wagner model [14]) with models that include the Pearce-Hall measure of associability, in which learning rates adjust dynamically based on the recent history of PEs [8]. These models incorporate two additional parameters, κ and η, which guide how EV and learning rate update in response to PEs (see Methods). We used Aikake's Information Criterion (AIC [15]) to evaluate goodness-of-fit while penalizing models with increased complexity (see Methods). The best fitting model was a hybrid model that assumes that error-driven learning is gated by associability and cue-evoked SCRs reflect the combined contribution of EV and associability (Table 1 and Figure 1), replicating previous work on feedback-driven aversive learning [5,6,11]. Parameters from this model were used to generate regressors for fMRI analyses of feedback-driven learning.

To understand potential effects on dynamic learning, we introduced a reversal parameter, *ρ*, that determines the extent to which EV and associability reverse when instructions are delivered [2,16]. If ρ = 1, the EV and associability of the CSs are swapped completely upon instruction, whereas if ρ = 0, each CS maintains its current EV and associability and the model reduces to the standard experiential one considered above (see Methods). We modified the aforementioned family of hybrid models to incorporate this additional parameter and fitted to the Instructed Group's SCRs (see [2], for visualization of SCRs over time in each group). The best model was the modified hybrid model that assumed SCR reflects the combination of associability and expected value – i.e. the same model that fit best in the Uninstructed Group, but with the additional instructed reversal parameter (Table 2 and Figure 1).

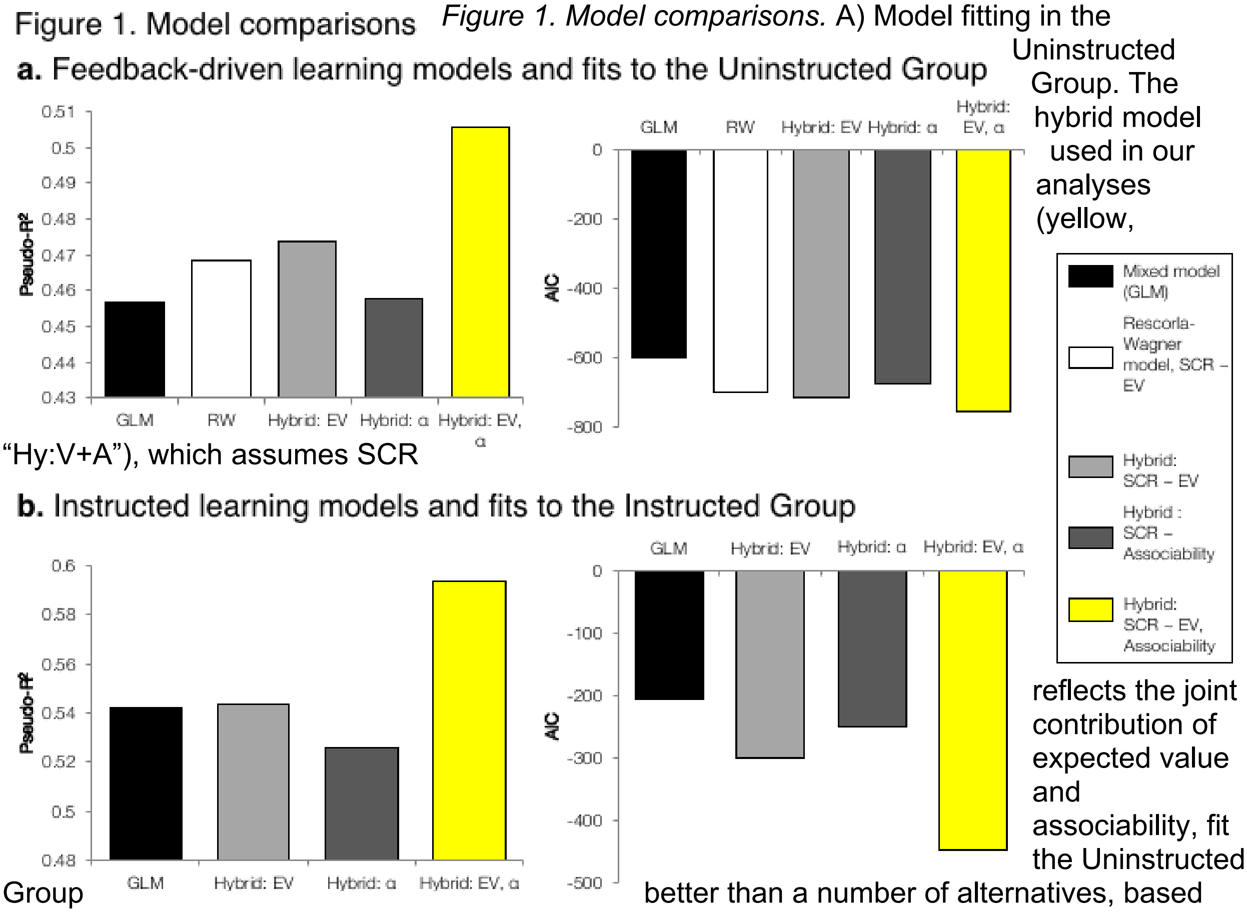

Figure 1. Model comparisons. A) Model fitting in the Uninstructed Group. The hybrid model used in our analyses (yellow, "Hy:V+A"), which assumes SCR reflects the joint contribution of expected value and associability, fit the Uninstructed Group better than a number of alternatives, based

both on absolute measures (pseudo-R^2, Left), as well as relative measures (Aikake's information criterion [AIC], Right). Fit quality is compared with a general linear model ("GLM") that captures the reversal of conditioned responses without dynamic learning (GLM reported in Table S1; black). The model also outperformed the standard Rescorla-Wagner model ("R-W") that assumes learning rate remains constant across time, as well as hybrid models of dynamic learning that assume SCR reflects either expected value ("Hy:V") or associability ("Hy:A"). B) The hybrid model also provided the best fits to the Instructed Group relative to the GLM or models that assumed SCR reflected either expected value or associability alone.

The best-fitting parameters when fit across subjects revealed that EV and associability reversed almost completely at the time of instructions ($\rho = 0.939$; Table 2), and EV did not update with PE ($\kappa = 0.0$), suggesting that instructions directly influence EV without further learning due to subsequent reinforcement, as illustrated in Figure 2. To check the robustness of the instructed learning model, we also fit it to SCRs within the Uninstructed Group (Table 2), for whom the additional effect should not be observed since no instructions were given. Parameter estimates were indeed consistent with associations not reversing at the time when the instructions would have been delivered ($\rho = 0.122$ from across-subjects model). As expected, instructed reversal parameters differed across groups, based on within-subject model fits (*Instructed* > *Uninstructed;* t(38) = 2.87, p < .01; see Methods and Table 2).

Figure 2. Instructions influence skin conductance responses during aversive learning.

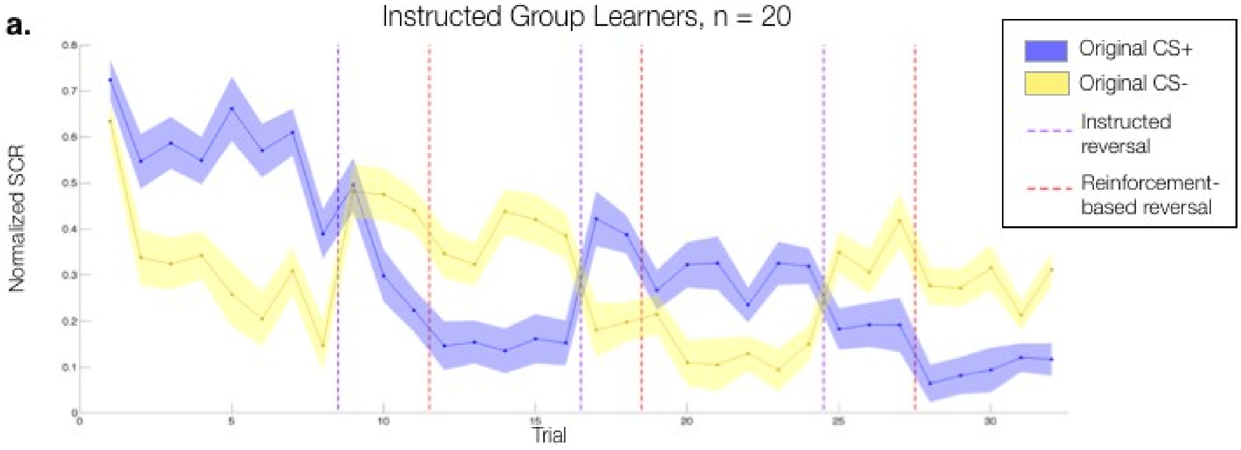

Figure 2. Instructions influence skin conductance responses during aversive learning. Within the Instructed Group, skin conductance responses increased immediately when participants were informed that the contingencies had reversed (purple dashed line). Reinforcing the new CS+/ previous CS- led to no additional increase in SCR (red dashed line). Shaded area reflects within-subjects error. Results for the Uninstructed Group and averages by phase are reported in [2].

Neural correlates of feedback-driven and instructed aversive learning. Our goal was to determine whether and how the neural systems that support feedback-driven aversive learning are modulated by instructions. Thus our computational neuroimaging analyses proceeded in two stages, guided by the two quantitative models described above. First, we used the hybrid model fit to behavior in the Uninstructed Group to generate the time course of learning from reinforcement alone. This allowed us to isolate the neural correlates of feedback-driven associability and feedback-driven PE in the Uninstructed Group, replicating previous results, and to test whether analogous responses were present in the Instructed Group despite the presence of instructions about contingencies and reversals. Second, turning our attention wholly to the Instructed Group, we used the modified hybrid model that was fit to that group to generate the time course of instructed learning and to isolate learning-related brain responses that update with instructions. We directly compared the two learning models within the amygdala and ventral striatum in the Instructed Group to determine whether these *a priori* regions of interest (ROIs) were sensitive to feedback-driven learning or whether they updated with instructions. Finally, we tested the conclusions from quantitative models with task-based analyses that relied strictly on our experimental design thus eliminating the influence of assumptions derived from our models.

Dissociable effects of instructions on feedback-driven learning in striatum and amygdala. We first focused on the neural correlates of feedback-driven associability and PE. Regressors were based on the best-fitting parameters from the hybrid model fit to the Uninstructed Group (see Methods). ROI-based and voxel-wise analyses within the Uninstructed

Group replicated previous findings [5], such that VS responses correlated with PE, while amygdala responses correlated with associability (see Figure 3 and Table 3). We then tested for correlates of the same feedback-driven regressors in the Instructed Group, and used ROI-based ANOVAs to test for group differences in feedback-driven learning (see Methods). We observed a dissociation in the effects of instructions on feedback-driven learning-related activity in these ROIs. There was a main effect of Group on feedback-driven PE signaling in VS (F(1, 38) = 7.04, p < .05), such that we observed striatal PEs in the Uninstructed Group, but not the Instructed Group (see Figure 3 and Table 3).

Figure 3. Effects of instructions on neural correlates of feedback-driven aversive learning

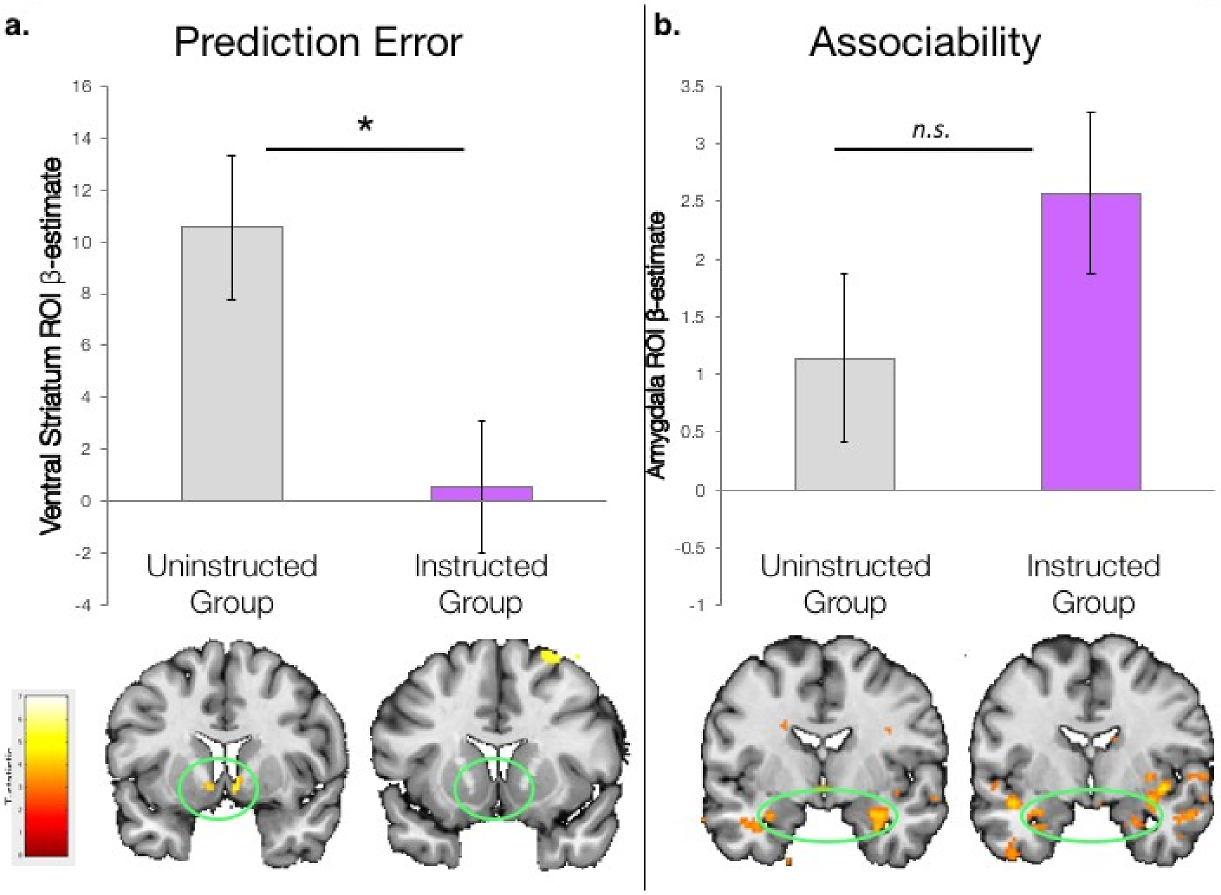

Figure 3. Effects of instructions on neural correlates of feedback-driven aversive learning. A) Neural correlates of feedback-driven prediction error (PE), based on fits to the Uninstructed Group. We found a significant difference in feedback-driven PE within our *a priori* ventral striatum (VS) ROI bilaterally (top), driven by significant striatal PE only in the Uninstructed

Group. ROI-based results were consistent with voxel-wise analyses (bottom). B) Neural correlates of feedback-driven associability, based on fits to the Uninstructed Group. Bilateral amygdala responses correlated with feedback-driven associability in both groups, based on both ROI-level and voxel-wise analyses (see also Table 6, Figures 1 and 4 and Tables 4 and 5).

However, there was no main effect of Group on feedback-driven associability in the amygdala (F(1, 38) = 2.0; p > .15), as the amygdala was correlated with feedback-driven associability in both groups (see Table 3), with stronger correlations in left amygdala (Laterality effect: F(1, 38) = 4.36, p < .05). Voxel-wise analyses confirmed these ROI-based dissociations: the Instructed Group showed feedback-driven associability bilaterally in the amygdala, but showed no evidence of feedback-driven PE in the VS (see Figure 3). The dissociation in feedback-driven learning was also apparent when we analyzed the entire sample, including those who did not exhibit measurable reversals in SCR (see Figure 4). Additional regions that tracked feedback-driven associability and PE across the task and/or showed group differences in these effects in whole brain analyses are reported in Figure 4 and Tables 4 and 5.

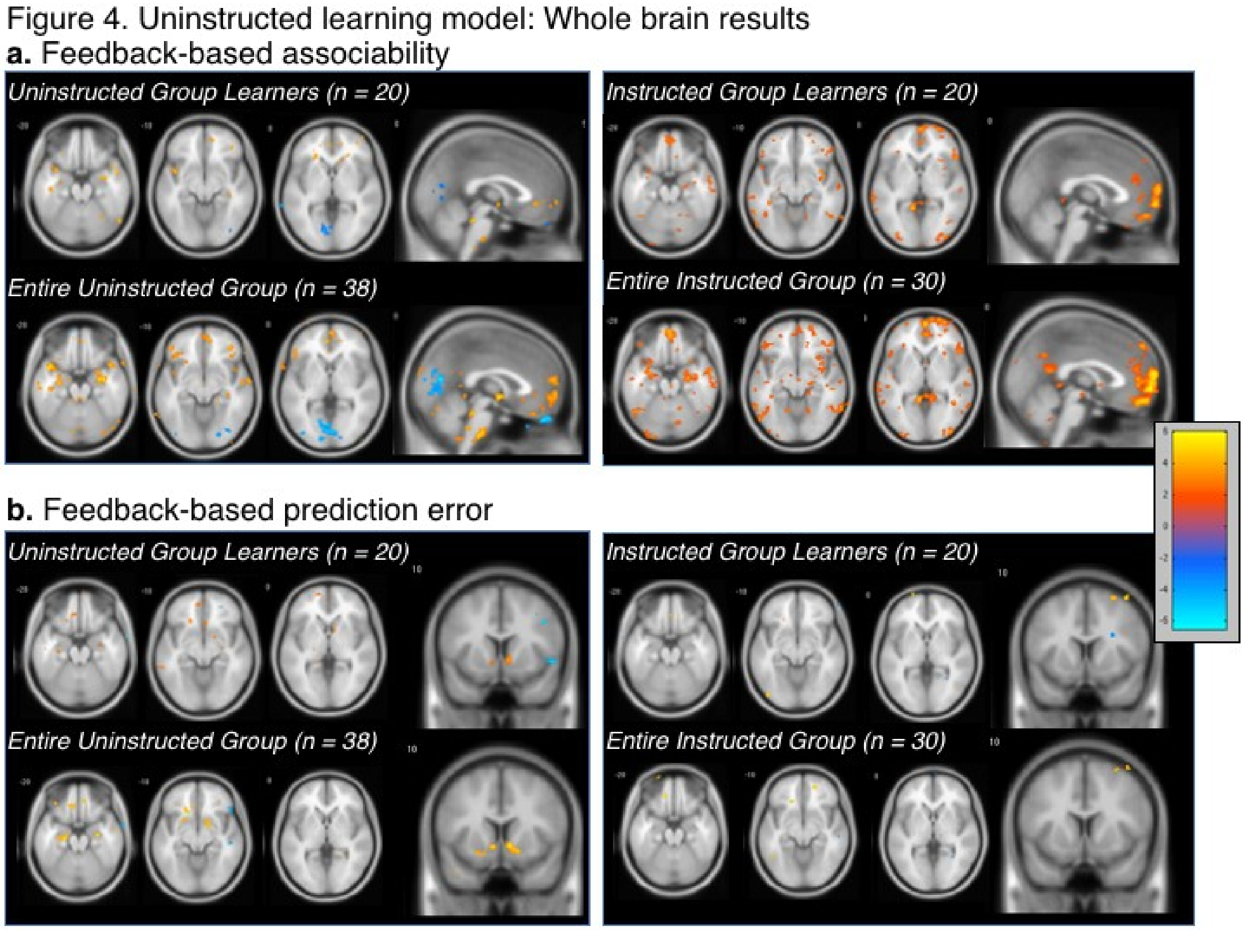
Figure 4. Uninstructed learning model: Whole brain results
a. Feedback-based associability
Uninstructed Group Learners (n = 20)
Instructed Group Learners (n = 20)
Entire Uninstructed Group (n = 38)
Entire Instructed Group (n = 30)
b. Feedback-based prediction error
Uninstructed Group Learners (n = 20)
Instructed Group Learners (n = 20)
Entire Uninstructed Group (n = 38)
Entire Instructed Group (n = 30)
6
4
2
0
-2
-4
-6

Figure 4. Uninstructed adaptive learning model: Whole-brain results. Whole-brain, voxel-wise results of feedback-driven learning in learners and non-learners. Parametric regressors were derived from the hybrid model fit to SCRs from learners in the Uninstructed Group (see Methods). Data are illustrated at a voxel-wise uncorrected threshold of $p < .001$, with an extent threshold of 10 voxels. Tables 4 and 5 present FWE-corrected results.

These results indicate that feedback-driven associability based on the Uninstructed Group's SCRs captured the dynamics of amygdala responses in both the Uninstructed Group and the Instructed Group, but striatal timecourses differed across groups. This would be expected if the striatum were sensitive to instructions, i.e. if EV (and thus PE) updated when individuals were instructed about contingencies and reversals, and if, conversely, associability related signaling in the amygdala was driven by reinforcement alone. To further investigate this interpretation, and to verify that seemingly feedback-driven activity in amygdala in the

Instructed Group is not better explained as arising from instruction-driven signals there, we turned to the modified hybrid model containing instructions.

Neural correlates of instructed aversive learning. We used parameters from the best-fitting instructed learning model to isolate the neural correlates within the Instructed Group of aversive learning that updates value and associability on the basis of instructions. We included EV in addition to associability and PE to capture directly the effect of instructions. EV updates immediately with instructions and does not update further with actual reinforcement, due to the low κ parameter. Voxel-wise analyses and ROI-based analyses revealed that VS and ventromedial prefrontal cortex/ medial orbitofrontal cortex (VMPFC/OFC) tracked instruction-based EV. We observed positive correlations with instruction-based EV in bilateral VS and a negative correlation in the VMPFC/OFC (see Figures 5 and 6 and Tables 6-8). We note that the robust bilateral caudate activation observed in voxel-wise analyses was not captured in ROI-wise analyses, as our *a priori* VS ROIs fell inferior to the focus of activation.

We also examined the neural correlates of associability and PE from the instructed learning model. We did not observe instruction-driven associability signals in our amygdala ROIs on a voxel-wise or ROI-wise basis. However, analyses did reveal instruction-driven PE coding in the striatum (Figure 5b; Table 6). Voxel-wise analyses revealed that the right VS ROI was correlated positively with instruction-driven PE (oriented with positive BOLD excursions for unexpected aversive events), and whole brain analyses revealed a negative correlation between instructed PE and activation in bilateral putamen. ROI-based analyses of VS did not show a reliable relationship with PE, presumably because of the mixed valence in the region. Whole brain exploratory results for additional regions that track instructed aversive learning are reported in Figure 6 and Tables 7 and 8.

Figure 5. Neural correlates of instructed aversive learning.

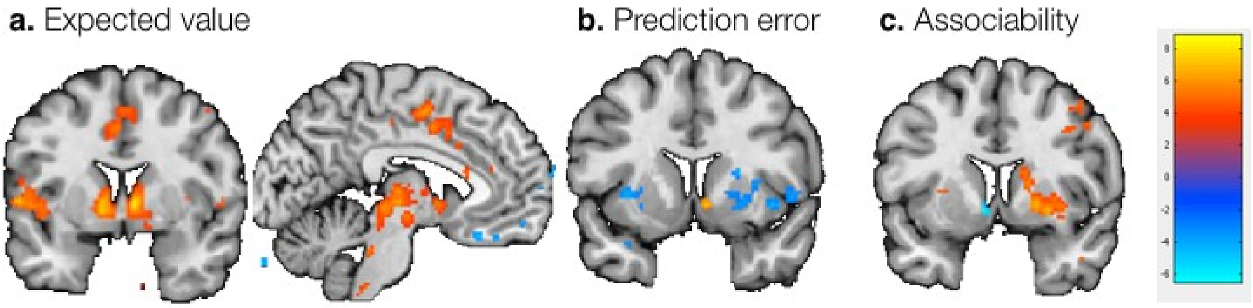

Figure 5. Neural correlates of instructed aversive learning. Neural correlates of dynamic error-driven learning from a model where expected value (EV) and associability are sensitive to instructions (see also Figure 6 and Tables 7 and 8). A) We observed positive correlations with EV in ventral striatum (VS; left). The VMPFC/OFC was negatively correlated. B) Our instructed model revealed positive instructed aversive PE signals (increases with unexpected aversive outcomes) in the right VS near the nucleus accumbens and negative PEs (increases with unexpected omissions of aversive shock, consistent with an appetitive PE) in bilateral putamen. C) Instructed associability was positively correlated with responses in right putamen and caudate, and negatively correlated with responses in left VS. We saw no relationship between amygdala and instructed associability.

Feedback-driven vs instructed learning in the Instructed Group. The preceding results, from separate analyses of signals driven by instructions or feedback, suggest that striatal (and VMPFC/OFC) responses are sensitive to instructions, whereas amygdala responses are driven by reinforcement. To directly and formally compare the neural correlates of instructed and uninstructed aversive learning in participants exposed to both forms of feedback, we performed statistical analyses of our two learning models in our striatal and amygdala ROIs. For each Instructed Group participant, beta coefficients were extracted from each ROI for each area's

Figure 6. Instructed learning model: Whole brain results

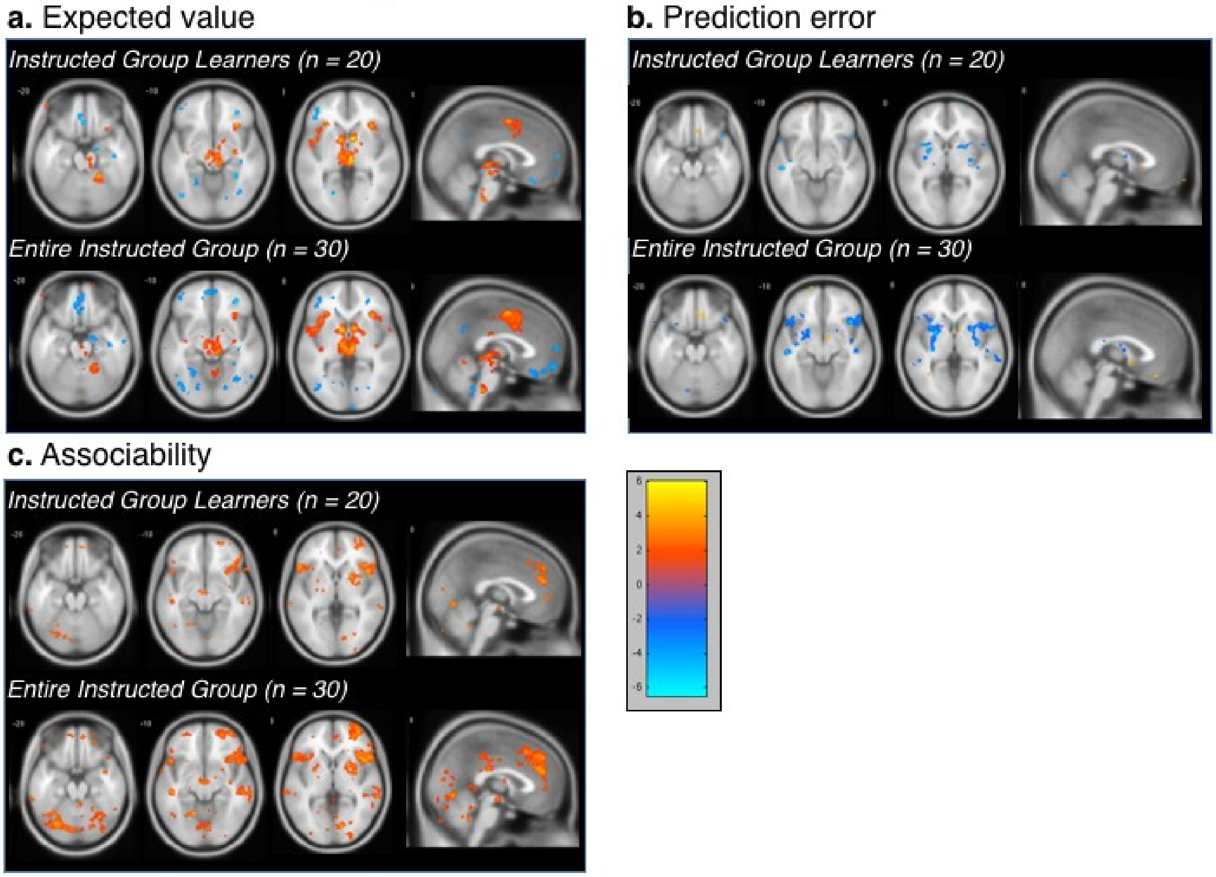
a. Expected value
Instructed Group Learners (n = 20)
Entire Instructed Group (n = 30)
b. Prediction error
Instructed Group Learners (n = 20)
Entire Instructed Group (n = 30)
c. Associability
Instructed Group Learners (n = 20)
Entire Instructed Group (n = 30)
-20
-10
0
0
6
4
2
0
-2
-4
-6

Figure 6. *Instructed adaptive learning model: Whole-brain results.* Whole-brain, voxel-wise results of the neural correlates of instructed learning in learners and non-learners within the Instructed Group. Parametric regressors were derived from the modified hybrid model fit to SCRs from learners in the Instructed Group (see Methods). Data are illustrated at a voxel-wise uncorrected threshold of $p < .001$, with an extent threshold of 10 voxels. Tables 7 and 8 present FWE-corrected results.

relevant learning-related signal (i.e. associability in amygdala, PE in VS) from both feedback-driven and instructed models. We used ANOVAs to evaluate effects of Model (instructed vs uninstructed) signal and Laterality. There was a significant effect of Model on the amygdala ($F(1, 19) = 7.95$, $p = .011$), such that amygdala responses were more related to Uninstructed associability than Instructed associability (*left amygdala:* $t(19) = 2.69$, $p < .05$, *right amygdala:* $t(19) = 2.41$, $p < .05$). We observed no effect of Model on striatal PE (all p's > .5) presumably due to the observed mixture of positive and negative instructed PE responses within the boundaries of our ROIs. Finally, we did not observe Laterality effects ins either ROI, nor any Model x Laterality interactions (all p's > .2).

Discussion

The aim of this experiment was to determine whether and how instructions modulate the neural mechanisms of adaptive aversive learning. Quantitative modeling in the Uninstructed Group revealed that the amygdala tracked associability and the striatum tracked prediction error (PE), replicating previous work [5]. This feedback-driven associability signal also explained the dynamics of amygdala responses in the Instructed Group, suggesting that the amygdala learned from reinforcement history, irrespective of instructions. In contrast to the amygdala, responses in the striatum and VMPFC/OFC tracked learning-related responses that updated immediately with instruction. The neural dissociations we observed here are consistent with the dissociations reported in our previous work that assumed learning rates were stable over time [2]. The present report builds on this work by demonstrating that a hybrid model provides a better fit to the data, and by revealing that prediction errors and expected value update based on instructed knowledge, while associability depends primarily on experienced outcomes. Below, we discuss the implications of these findings, their relationship to previous work, and important outstanding questions.

We found evidence for aversive PEs (increased activation to unexpected aversive outcomes) in VS in our Uninstructed Group, which replicates previous work [5,17,18]. However, when participants were instructed about contingencies, the VS no longer correlated with feedback-driven PEs. Instead, the striatum tracked parameters that updated with instructions. We found that the caudate tracked EV, which updated with instructions and remained constant over the duration of each run, while the VS and putamen tracked PE, which varied over time as a function of the intermittent reinforcements, relative to instruction-based EV. These quantitative findings were confirmed with task-based analyses that revealed that differential responses in the striatum reversed immediately upon instruction. Analyses also revealed immediate effects of instruction on the VMPFC/OFC, which has been linked with expected value and reversal learning in both appetitive and aversive domains [12,19,20]. The VMPFC/OFC tracked instructed expected value that reversed upon instruction and did not update further with aversive feedback, perhaps consistent with the view that this region encodes task state [21].

While findings in striatum and VMPFC/OFC replicate and extend work on instructed reward learning, findings in the amygdala reveal an important distinction in this new analysis of adaptive learning. The amygdala responded dynamically to threats in the environment similarly in both groups and showed no evidence for updating with instructions, consistent with our previous findings using a fixed learning rate and a standard Rescorla-Wagner model [2].

Associability signals in the amygdala replicate previous work on aversive reversal learning [5], and add to a growing literature on adaptive learning across domains[7,10,22]. Associability – and hence amygdala responding – is highest when attention is drawn by unexpected events in the environment that are likely to shape subsequent learning, i.e. at the start of learning, and when a shock is paired with a stimulus that was thought to be safe. The finding that neural correlates of EV and PEs are susceptible to verbal instructions, whereas associability signals in the amygdala are not, may relate to the different computational roles of these variables. In particular, computationally, associability is thought to correspond to uncertainty about a stimulus's value [22]. Thus, even if value can come under control of instructions, amygdala-dependent uncertainty about that value may be strictly experiential and implicit.

In conclusion, reanalysis of our previously published data using a hybrid model of adaptive learning builds on the dissociations revealed with a parsimonious model that assumed that learning rates are constant over time. Value coding and prediction error in corticostriatal circuits updated immediately with instructions. However, associability in the amygdala required aversive feedback in order to update. Our findings also contribute to a growing body of work on adaptive learning and reveal that cognitive knowledge adds another layer of complexity to the neural coding of value and expectation. Future work should directly compare appetitive and aversive learning to understand whether associability acts similarly in guiding adaptive reward learning, and how such learning is modulated by instruction.

Methods

This report includes a reanalysis of behavioral and neural data from [2]. In brief, sixty-eight participants completed an aversive reversal learning task in which a CS+ (angry face) was paired with an electric shock (US) with a 30% reinforcement rate, while a CS- (a different angry face) was not paired with a shock. Contingencies reversed three times during the experiment. An Instructed Group (N = 30) was informed about contingencies and prior to reversals, while an Uninstructed Group learned through experience. Please see [2] for complete details on participants and experimental procedures.

Quantitative modeling.

Our learning models assume that SCR correlates with dynamic quantities derived from feedback-driven or instructed adaptive learning models. Below we describe the quantitative models we evaluated, followed by our general procedures for model fitting and model comparison.

Feedback-driven model. We fit several learning models to trial-by-trial SCRs from the Uninstructed Group to test whether reinforcement learning models can explain fear-conditioning behavior in the context of multiple reversals, and to generate predictors that will isolate the brain mechanisms of feedback-driven learning irrespective of instructed knowledge. The standard Rescorla-Wagner model learns an expected value (EV), denoted as V, for each CS, x, and assumes a fixed learning rate (α):

(1) $V_{n+1}(x_n) = V_n(x_n) + \alpha\delta_n$

(2) $\delta_n = r_n - V_n(x_n)$

(where $r = 1$ for shock, $r = 0$ for no shock, n denotes the current trial, and δ = prediction error (PE), and V = EV). To derive the best fits for the Rescorla-Wagner model, we set α and V_0 as free parameters, and evaluated regressions that assumed trial-by-trial SCRs correlate with EV (see "General procedures for model fitting and model comparison", below). Hybrid models were based on the Pearce-Hall model, wherein learning is gated dynamically depending on two free parameters, κ and η:

(3) $V_{n+1}(x_n) = V_n(x_n) + \kappa\alpha_n(x_n)\delta_n$

(4) $\alpha_{n+1}(x_n) =\eta|\delta_n| + (1-\eta)\alpha_n(x_n)$,

We compared three instantiations of the hybrid model that varied as a function of which parameters were fit to SCR. Model fits were based on regressions that assumed SCRs correlate with trial-by-trial estimates of either a) expected value, b) associability, or c) the joint contribution of value and associability [5]. To limit free parameters in the hybrid models, we assumed that learning would be maximal at the start of the task and set α_0 to 1.0 [5,11]. We also set the expected value for both CSs at $V_0 = 0.5$, as participants in the Uninstructed Group had no information about cue contingencies.

Instructed model. To account for the influence of instructions, we incorporated an additional reversal parameter (ρ) that determines the extent to which expected value and associability reverse at the time when instructions are delivered [2,16]. Learning proceeded as in the uninstructed learning model above (Eqs. 3 & 4) in all cases until instructions were delivered (i.e. immediately following trials 20, 40, and 60). At the time of instructions, for each of the two cues (x_a and x_b), EV was computed as the sum of the current cue's value multiplied by 1-ρ, plus the other cue's value multiplied by ρ:

(5) $V_{n+1}(x_a) = \rho * V_n(x_b) + (1-\rho) * V_n(x_a)$

(6) $V_{n+1}(x_b) = \rho * V_n(x_a) + (1-\rho) * V_n(x_b)$

Thus if $\rho = 0$, each cue retains its value, whereas if $\rho = 1$, cue x_a acquires the value of cue x_b. Associability was also assumed to reverse upon instruction according to the same parameter:

(7) $\alpha_{n+1}(x_a) = \rho * \alpha_n(x_b) + (1-\rho) * \alpha_n(x_a)$

(8) $\alpha_{n+1}(x_b) = \rho * \alpha_n(x_a) + (1-\rho) * \alpha_n(x_b)$.

Learning then proceeded according to the feedback-driven model until the next instructions were delivered. As with the feedback-driven model above, we assumed learning would be maximal at the start of the task, and assumed an initial α of 1.0. Because the Instructed Group was informed about the original cue contingencies, we assumed asymmetrical initial expected values (V_0(CS+) = .75, V_0(CS-) = .25). κ , η, and ρ were modeled as free parameters.

General procedures for model fitting and model comparison. As described in Atlas [2], models were fit both across subjects (with learning parameters modeled as fixed parameters with varying slopes) and within subjects. Complete details of model fitting and model comparison are reported in [2]. For each iteration of model fitting (i.e. test of a given parameter value) in Matlab's fminsearch.m program, candidate parameters were applied to each individual subject's trial order to generate a predicted timecourse (value, associability, or both, depending on the model). For each participant, the resulting regressor(s) was combined with an overall intercept as well as linear nuisance regressor to capture habituation.

Parameter estimation and model comparison. We used maximum likelihood estimation and Matlab's fminsearch function to determine best-fitting parameters in all models. Model comparisons in across-subjects fits were based on Akaike's Information Criterion (AIC; [15]), which evaluates log likelihood and penalizes models with additional parameters. AIC was computed as:

[illegible],

where *n* denotes the number of parameters, N denotes the number of observations, and SSE refers to the sum squared error or deviance of the residuals. For the mixed-effect fits in which the learning model parameters were fixed within each group, we compared models based on a single aggregate AIC score for the fit over all subjects. To identify the best-fitting model in the fits to each individual separately (within-subjects analyses), we submitted AIC scores for each individual to SPM8's Bayesian model selection routine (spm_BMS; [23]), which compares candidate models under the assumption that the true model may vary from subject to subject as a random effect. We used spm_BMS's exceedance probabilities (the estimated probability that each model is the most common one in the population) to determine the most likely model.

FMRI data acquisition and analysis. Please see [2] for complete details on FMRI data acquisition and processing.

General procedures for neuroimaging analyses. We conducted two whole-brain analyses using SPM8: 1) An analysis of feedback-driven aversive learning in all subjects, based on regressors from the hybrid model fitted to SCRs from the Uninstructed Group; and 2) An analysis of instructed aversive learning in the Instructed Group, based on regressors from the instructed learning model fitted to SCRs from the Instructed Group. First-level analyses employed the general linear model (GLM) in SPM8 without default implicit thresholding. All events were convolved with a canonical gamma-variate hemodynamic response function (HRF).

Group results were obtained using SPM8's standard summary statistics approach, with one-sample t-tests for within-group analyses and independent t-tests for comparisons across groups. We also performed ROI-wise t-tests and ANOVAs using standard functions (ttest.m, ttest2.m, and anovan.m) in Matlab. ANOVAs focusing on the feedback-driven learning model assess main effects of Group (Instructed vs. Uninstructed) and Laterality. ANOVAs focusing on the Instructed Group assess main effects of Model (instructed learning vs. feedback-driven learning) and Laterality. Post-hoc t-tests are reported in Tables 3 and 6.

Our main analyses focus on results in our *a priori* ROIs: amygdala, ventral striatum, and VMPFC/OFC. Amygdala and striatum ROIs were defined based on the MNI template, and are available at http://wagerlab.colorado.edu/tools. Additional details on ROI selection and creation are available in [2]. We used custom Matlab code (available at http://wagerlab.colorado.edu/tools) to extract and average across ROI-wise data, and we report ROI-wise analyses at standard $p < .05$. Voxel-wise ROI-based results were obtained through SPM8, and we report voxel-wise small-volume cluster-corrected results (initial threshold $p < .001$, FWE $p<.05$), unless otherwise noted. We also present whole-brain exploratory analyses in Tables 4, 5, 7 and 8 using whole-brain cluster-corrected thresholds (FWE $p < .05$). We report coordinates in Montreal Neurological Institute (MNI) space. Anatomical labels are based on the SPM anatomy toolbox [24]. Whole brain results are visualized at voxel-wise uncorrected thresholds ($p<.001$, $k = 5$).

Neural correlates of feedback-driven and instructed aversive learning. Regressors for neural analyses of learning-related signals were derived from the across-subjects fits to SCR, which generated best-fitting parameters (κ and η in both models; ρ in the instructed learning model). Our design had two trial orders, so we applied the best-fitting parameters to each sequence of stimuli to generate regressors for PE, EV, and associability that were specific to each trial order. We used parameters fit across subjects in the Uninstructed Group to generate regressors for each trial order that were used to identify neural correlates of feedback-driven learning in all participants. Thus, the feedback-driven learning analyses use identical regressors for both the Uninstructed Group and Instructed Group, varying only as a function of trial order. EV was excluded from this analysis due to collinearity stemming from its algebraic relationship with PE, but we note that conclusions about amygdala associability are consistent whether or not EV is included in whole-brain analyses (striatal prediction error is reduced when EV is incorporated in the uninstructed model, as expected given the algebraic relationship between EV, PE, and the trial feedback, which is also included in the model). Analyses focusing on the neural correlates of instructed aversive learning were based on parameters from the instructed learning model fit across participants in the Instructed Group.

Our learning analyses modeled cue onset and cue offset as two discrete events, with separate parametric modulators. Consistent with previous work [5], three regressors modulated the cue offset event: 1) shock occurrence (0 for trials with no shock, 1 for shock trials), 2) associability, and 3) PE. EV was included as a parametric modulator at the time of cue onset in the instructed learning model.

Acknowledgments

We would like to thank Eyal Bar-David, Rachel Mojdehbakhsh, and Keith Sanzenbach for assistance with fMRI scanning, and Manurhadi Perera, Augustus Baker, and Erin Furlong for assistance with SCR data processing and subject recruitment. This work was supported by an NIH grant awarded to E. A. Phelps (RO1MH097085), and by the Intramural Research program of the NIH's National Center for Complementary and Integrative Health._

References

[1] P. Dayan, B.W. Balleine, Reward, motivation, and reinforcement learning., Neuron. 36 (2002) 285–298. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=12383782&retmode=ref&cmd=prlinks.

[2] L.Y. Atlas, B.B. Doll, J. Li, N.D. Daw, E.A. Phelps, Instructed knowledge shapes feedback- driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala, Elife. 5 (2016). doi:10.7554/eLife.15192.

[3] B.B. Doll, W.J. Jacobs, A.G. Sanfey, M.J. Frank, Instructional control of reinforcement learning: A behavioral and neurocomputational investigation, Brain Res. 1299 (2009) 74–94. doi:10.1016/j.brainres.2009.07.007.

[4] J. Li, M.R. Delgado, E. Phelps, How instructed knowledge modulates the neural systems of reward learning, Proc Natl Acad Sci U S A. 108 (2011) 55–60. doi:10.1073/pnas.1014938108/-/DCSupplemental.

[5] J. Li, D. Schiller, G. Schoenbaum, E. Phelps, N.D. Daw, Differential roles of human striatum and amygdala in associative learning, 14 (2011) 1250–1252. doi:10.1038/nn.2904.

[6] S. Zhang, H. Mano, G. Ganesh, T. Robbins, B. Seymour, Dissociable Learning Processes Underlie Human Pain Conditioning, Curr. Biol. (2016). doi:10.1016/j.cub.2015.10.066.

[7] T.E.J. Behrens, M.W. Woolrich, M.E. Walton, M.F.S. Rushworth, Learning the value of information in an uncertain world, Nat. Neurosci. 10 (2007) 1214. https://doi.org/10.1038/nn1954.

[8] J.M. Pearce, G. Hall, A Model for Pavlovian Learning : Variations in the Effectiveness of

Conditioned But Not of Unconditioned Stimuli, 87 (1980) 532–552.

[9] P. Dayan, N.D. Daw, Decision theory, reinforcement learning, and the brain, Cogn. Affect. &amp; Behav. Neurosci. 8 (2008) 429–453. doi:10.3758/CABN.8.4.429.

[10] M.R. Nassar, R.C. Wilson, B. Heasly, J.I. Gold, An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment, J. Neurosci. 30 (2010) 12366–12378. doi:10.1523/JNEUROSCI.0822-10.2010.

[11] S. Boll, M. Gamer, S. Gluth, J. Finsterbusch, C. Büchel, Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans, Eur. J. Neurosci. 37 (2012) 758–767. doi:10.1111/ejn.12094.

[12] D. Schiller, I. Levy, Y. Niv, J.E. Ledoux, E. Phelps, From Fear to Safety and Back: Reversal of Fear in the Human Brain, J. Neurosci. 28 (2008) 11517–11525. doi:10.1523/JNEUROSCI.2265-08.2008.

[13] B.B. Doll, K.E. Hutchison, M.J. Frank, Dopaminergic Genes Predict Individual Differences in Susceptibility to Confirmation Bias, J. Neurosci. 31 (2011) 6188–6198. doi:10.1523/JNEUROSCI.6486-10.2011.

[14] R.A. Rescorla, A.R. Wagner, A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcment, in: A. Black, W. Prokasky (Eds.), Class. Cond. II Curr. Res. Theory, Appleton-Century-Crofts, New York, 1972: pp. 64–99. papers3://publication/uuid/81E7E5F0-BC49-4FB7-B0A8-FAD0F56F17E0.

[15] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (1974) 716–723. doi:10.1109/TAC.1974.1100705.

[16] L.Y. Atlas, E.A. Phelps, Prepared stimuli enhance aversive learning without weakening the impact of verbal instructions, Learn. Mem. 25 (2018). doi:10.1101/lm.046359.117.

[17] M.R. Delgado, J. Li, D. Schiller, E. Phelps, The role of the striatum in aversive learning and aversive prediction errors, Philos. Trans. R. Soc. B Biol. Sci. 363 (2008) 3787–3800. doi:10.1098/rstb.2008.0161.

[18] B. Seymour, J.P.O. Doherty, P. Dayan, M. Koltzenburg, A.K. Jones, R.J. Dolan, K.J. Friston, R.S. Frackowiak, Temporal difference models describe higher-order learning in humans, Nature. 429 (2004) 664–667. doi:10.1038/nature02636.1.

[19] E.A. Murray, J.P. O'Doherty, G. Schoenbaum, J.P. O&apos;Doherty, G. Schoenbaum, What We Know and Do Not Know about the Functions of the Orbitofrontal Cortex after 20 Years of Cross-Species Studies, J. Neurosci. 27 (2007) 8166–8169. doi:10.1523/JNEUROSCI.1556-07.2007.

[20] P.H. Rudebeck, R.C. Saunders, A.T. Prescott, L.S. Chau, E.A. Murray, Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating, Nat. Publ. Gr. (2013) 1–8. doi:10.1038/nn.3440.

[21] R.C. Wilson, Y.K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal Cortex as a Cognitive Map of Task Space, Neuron. 81 (2014) 267–279. doi:10.1016/j.neuron.2013.11.005.

[22] P. Dayan, S. Kakade, P.R. Montague, Learning and selective attention., Nat. Neurosci. 3 Suppl (2000) 1218–1223. doi:10.1038/81504.

[23] K.E. Stephan, W.D. Penny, J. Daunizeau, R.J. Moran, K.J. Friston, NeuroImage Bayesian model selection for group studies, Neuroimage. 46 (2009) 1004–1017. doi:10.1016/j.neuroimage.2009.03.025.

[24] S.B. Eickhoff, The Human Parietal Operculum. II. Stereotaxic Maps and Correlation with Functional Imaging Results, Cereb. Cortex. 16 (2005) 268–279.

doi:10.1093/cercor/bhi106.

Figure Legends

Figure 1. Model comparisons. A) Model fitting in the Uninstructed Group. The hybrid model used in our analyses (yellow, “Hy:V+A”), which assumes SCR reflects the joint contribution of expected value and associability, fit the Uninstructed Group better than a number of alternatives, based both on absolute measures (pseudo-R^2, Left), as well as relative measures (Aikake’s information criterion [AIC], Right). Fit quality is compared with a general linear model (“GLM”) that captures the reversal of conditioned responses without dynamic learning (GLM reported in Table S1; black). The model also outperformed the standard Rescorla-Wagner model ("R-W") that assumes learning rate remains constant across time, as well as hybrid models of dynamic learning that assume SCR reflects either expected value (“Hy:V”) or associability (“Hy:A”). B) The hybrid model also provided the best fits to the Instructed Group relative to the GLM or models that assumed SCR reflected either expected value or associability alone.

Figure 2. Instructions influence skin conductance responses during aversive learning. Within the Instructed Group, skin conductance responses increased immediately when participants were informed that the contingencies had reversed (purple dashed line). Reinforcing the new CS+/ previous CS- led to no additional increase in SCR (red dashed line). Shaded area reflects within-subjects error. Results for the Uninstructed Group and averages by phase are reported in [2].

Figure 3. Effects of instructions on neural correlates of feedback-driven aversive learning. A) Neural correlates of feedback-driven prediction error (PE), based on fits to the Uninstructed Group. We found a significant difference in feedback-driven PE within our *a priori* ventral striatum (VS) ROI bilaterally (top), driven by significant striatal PE only in the Uninstructed Group. ROI-based results were consistent with voxel-wise analyses (bottom). B) Neural correlates of feedback-driven associability, based on fits to the Uninstructed Group. Bilateral amygdala responses correlated with feedback-driven associability in both groups, based on both ROI-level and voxel-wise analyses (see also Table 6, Figures 1 and 4 and Tables 4 and 5).

Figure 4. Uninstructed adaptive learning model: Whole-brain results. Whole-brain, voxel-wise results of feedback-driven learning in learners and non-learners. Parametric regressors were derived from the hybrid model fit to SCRs from learners in the Uninstructed Group(see Methods). Data are illustrated at a voxel-wise uncorrected threshold of $p < .001$, with an extent threshold of 10 voxels. Tables 4 and 5 present FWE-corrected results.

Figure 5. Neural correlates of instructed aversive learning. Neural correlates of dynamic error-driven learning from a model where expected value (EV) and associability are sensitive to instructions (see also Figure 6 and Tables 7 and 8). A) We observed positive correlations with EV in ventral striatum (VS; left). The VMPFC/OFC was negatively correlated. B) Our instructed model revealed positive instructed aversive PE signals (increases with unexpected aversive outcomes) in the right VS near the nucleus accumbens and negative PEs (increases with unexpected omissions of aversive shock, consistent with an appetitive PE) in bilateral putamen. C) Instructed associability was positively correlated with responses in right putamen and caudate, and negatively correlated with responses in left VS. We saw no relationship between amygdala and instructed associability.

Figure 6. *Instructed adaptive learning model: Whole-brain results.* Whole-brain, voxel-wise results of the neural correlates of instructed learning in learners and non-learners within the Instructed Group. Parametric regressors were derived from the modified hybrid model fit to SCRs from learners in the Instructed Group (see Methods). Data are illustrated at a voxel-wise uncorrected threshold of $p < .001$, with an extent threshold of 10 voxels. Tables 7 and 8 present FWE-corrected results.

Table 1. Feedback-driven learning: Model fitting in the Uninstructed Group (learners only, n = 20).[a]

	Model	Alpha	V_0	κ	η	ResErr	AIC
Across-subjects fits	Basic RL, SCR ~ Value	0.0489	0.1814	n/a	n/a	39.1685	-821.2532
	Associability, SCR ~ Value; fixed v0=.5 a0=1	n/a	n/a	0.2873	0.1818	38.7731	-834.2193
	Associability, SCR ~ Associability; fixed v0=.5 a0=1	n/a	n/a	1.0000	.0862	39.9753	-795.1958
	Associability, SCR ~ Associability + Value; fixed v0=.5 a0=1	n/a	n/a	0.1376	0.1578	36.4279	-911.9562
Within-subjects fits	Basic RL, SCR ~ Value	*M* =0.157, *SD* = 0.29	*M* = 0.518, *SD* = 0.42	n/a	n/a	n/a	-846.9776
	Associability, SCR ~ Value; fixed v0=.5 a0=1	n/a	n/a	*M* = 0.345, *SD* = 0.34	*M* =0.362, *SD* =0.38	n/a	-827.4531
	Associability, SCR ~ Associability; fixed v0=.5 a0=1	n/a	n/a	*M* =0.424, *SD* = .46	*M* =0.268, *SD* = 0.34	n/a	-842.6490
	Associability, SCR ~ Associability + Value; fixed v0=.5 a0=1	n/a	n/a	*M* = 0.304; *SD* = 0.37	*M* = 0.317, *SD* = 0.30	n/a	-923.2013

[a]. This table presents the results of quantitative learning models fit to skin conductance responses (SCR) from learners in the Uninstructed Group. We report best fitting parameters within each model, based on fminsearch in Matlab (see Methods). Res err = residual error from the mixed-effects ("across-subjects") model fits.

Table 2: Instructed model fitting.[b]

	Model	κ	η	ρ	Reserr	AIC
Across-subjects fits: Instructed Group (learners only, n = 20)	Associability, SCR ~ Value; a0=1, V0 = .75, .25	0.0310	0.0820	1.0000	53.7172	-251.2690
	Associability, SCR ~ Associability; a0=1, V0 = .75, .25	0.0353	0.7159	1.0000	55.8124	-300.2471
	Associability, SCR ~ Associability + Value; a0=1, V0 = .75, .25	0.0000	0.4992	0.9392	47.7620	-448.6511
Within-subjects fits: Instructed Group (learners only, n = 20)	Associability, SCR ~ Value; a0=1, V0 = .75, .25	*M* = 0.668, *SD* = 0.40	*M* = .496, *SD* = 0.02	*M* = .508, *SD* =0.03		-272.1274
	Associability, SCR ~ Associability; a0=1, V0 = .75, .25	*M* = 0.826, *SD* = 0.31	*M* = 0.501, *SD* = 0.07	*M* = .490 *SD* = 0.04		-183.0228
	Associability, SCR ~ Associability + Value; a0=1, V0 = .75, .25	*M* = 0.228, *SD* = .08	*M* = .325, *SD* = .08	*M* = .689, *SD* = .06		-665.2614
Across-subjects fits: Uninstructed Group (learners only, n = 20)	Associability, SCR ~ Associability + Value; a_0=1, V_0 = .5	0.0449	0.3219	0.1223	36.2247	-799.1048
Within-subjects fits: Uninstructed Group (learners only, n = 20)	Associability, SCR ~ Associability + Value; a_0=1, V_0 = .5	*M* = .298 *SE* = .08	*M* = .322; *SE* = .07	*M* = .379, *SE* = .09		-927.3583

[b]. This table presents the results of modified learning models that included a reversal parameter (ρ) that determined the extent to which expected value and associability exchange current estimates upon instructed reversal. Values of ρ = 1 denote a complete swap, whereas values of ρ = 0 denote that each CS retains its current value. Full methods are reported in Methods.

Table 3. Feedback-driven adaptive learning in regions of interest (ROIs).[c]

		Feedback-driven Associability		Feedback-driven Prediction Error	
		Left amygdala	Right amygdala	Left Ventral Striatum	Right Ventral Striatum
ROI-based analyses	*Uninstructed Group*	t(19)= 2.08⊥	*n.s.*	t(19)=3.73**	t(19)=3.06**
	Instructed Group	t(19)= 3.78**	t(19)=3.06**	*n.s.*	*n.s.*
	Uninstructed Group > Instructed Group	*n.s.*	*n.s.*	t(38)=2.44*	t(38)=2.24*
Voxel-based, small-volume-corrected xyz coordinates	*Uninstructed Group*	[-34 6 -22][b]	[30 -4 -20]	[-10 14 -6]	[8 10 -6]
	Instructed Group	[-20 0 -24]; [-34 -2 -20]	*n.s.*	*n.s.*	n.s.
	Uninstructed Group > Instructed Group	*n.s.*	*n.s.*	[-10 14 -6][c]	[6 8 -8][c]

[c] This table presents region-of-interest (ROI)-based results for learning parameters derived from the feedback-driven model fit to behavior in the Uninstructed Group. Within-group results refer to one-sample t-tests, and between groups results are based on independent-samples T-tests. ROI-based analyses report results of post-hoc t-tests to accompany the ANOVAs of feedback-driven learning reported in the main manuscript. Data are extracted and averaged across ROIs, and significance is evaluated at p < .05. Voxel-based analyses report small-volume corrected results for ROIs at p(FWE-corrected) < .05 unless otherwise noted.

** *p* < .01

* *p* < .05

⊥ one-tailed *p < .05*

[b] FWE-corrected p = .058

[c] Significant at voxel-wise uncorrected *p* < .005

Table 4, Neural correlates of feedback-driven learning in learners (n = 40).[d]

Learning Parameter	Contrast	Region	xyz	Cluster p (FWE-cor)	Cluster equivk	Peak equivZ
Feedback-driven associability	*Main effect across groups, Positive*	L Mid Orbital Gyrus; R Rectal Gyrus; R Inferior Frontal Gyrus (IFG) p. Orbitalis (VMPFC/MPFC/Frontal Pole)	0 64 -4	0	1940	5.44
		L Precuneus; L Cerebellum	-8 -54 16	0	387	5.19
		R Middle Temporal Gyrus; R Superior Temporal Gyrus	52 -12 -14	0	175	5.1
		R Middle Insula	42 -2 -6	0.006	83	5.09
		L IFG p. Orbitalis; L Insula	-30 34 -6	0	120	5.09
		R Cerebellum	16 -92 -32	0	206	4.88
		Posterior Hippocampus	38 -36 -4	0.022	66	4.84
		Precuneus (Bilateral)	6 -60 32	0.01	76	4.79
		L Amygdala; L Anterior Insula; L Inferior Temporal Gyrus	-36 6 -18	0	495	4.77
		R Amygdala; R Inferior Temporal Gyrus	36 -6 -18	0	216	4.7
		R Cerebellum; L Superior Temporal Gyrus	40 -52 -26	0	154	4.58
		Cerebellum (bilateral)	48 -74 -26	0.001	104	4.33
		Cuneus (V1)	4 -98 20	0.01	76	4.28
		L Middle Temporal Gyrus	-66 -28 -2	0	124	4.27
		L IFG p. Orbitalis (latOFC)	-42 38 -14	0.001	111	4.05
		L Middle Temporal Gyrus; L Inferior Temporal Gyrus	-54 -48 -6	0.026	64	3.87
	Main effect across groups, Negative	No clusters survive FWE-correction.				
	Uninstructed > Instructed	No clusters survive FWE-correction.				
	Instructed > Uninstructed	R Supramarginal Gyrus	52 -40 32	0	118	5.35
		R Middle Frontal Gyrus; R IFG p. Triangularis (latPFC)	50 30 2	0.001	105	4.44
		L Lingual Gyrus (V1)	-18 -76 -4	0	132	4.37
		L Middle Occipital Gyrus	-24 -88 12	0.002	96	4.37
		R Superior medial gyrus	10 70 2	0.037	60	4.07
	Uninstructed Group, Positive	L Amygdala; L Insula; L Temporal Pole	-42 4 -10	0	122	4.21
	Uninstructed Group, Negative	L Lingual Gyrus (V1)	-10 -76 2	0	126	4.48
	Instructed Group, Positive	R Superior Medial Gyrus; L Mid Orbital Gyrus (MPFC)	8 70 12	0	608	5.57
		L Lingual Gyrus; Cerebellar Vermis	-10 -48 0	0.004	88	5.12
		L Hippocampus; L Superior Temporal Gyrus	-42 -6 -14	0.003	94	4.97
		R Middle Occipital Gyrus; R Calcarine Gyrus	32 -94 8	0	198	4.86
		R Middle Frontal Gyrus; R Superior Frontal Gyrus (DLPFC)	34 64 6	0	430	4.78
		R Middle Temporal Gyrus	52 -12 -14	0.04	59	4.74
		L Precuneus; L Calcarine Gyrus	-8 -54 18	0	120	4.68
		L Rectal Gyrus (VMPFC/mOFC/Frontal Pole)	-2 54 -18	0	132	4.66
		R Anterior Insula	36 8 -8	0.043	58	4.65
		R IFG p. Triangularis; R IFG p. Orbitalis (BA45/ latOFC/ latPFC)	52 24 2	0	264	4.65
		R Middle Insula; R Hippocampus	42 -2 -6	0.001	108	4.62
		L Cerebellum	-38 -56 -24	0.007	80	4.54
		R Cerebellum	18 -92 -32	0.029	63	4.41
		R Superior Temporal Gyrus	62 -18 -6	0.024	65	4.32
		R Middle Temporal Gyrus	56 -40 6	0.014	72	4.31
		R Middle Frontal Gyrus (DLPFC)	30 34 42	0.01	76	4.27
		L ACC / MPFC / raCC	-4 50 8	0.004	87	4.18
		L Middle Temporal Gyrus	-70 -32 2	0.005	84	4.14
		R Cerebellum (Crus 1)	48 -76 -26	0.04	59	4
		L Middle Occipital Gyrus; L Inferior Occipital Gyrus	-34 -88 -2	0.001	106	3.96
		L Middle Temporal Gyrus	-60 -52 -10	0.04	59	3.95
		L Middle Frontal Gyrus (DLPFC, latPFC)	-38 58 14	0.005	84	3.91
		R Middle Orbital Gyrus	26 54 -14	0.047	57	3.79
		ACC (DMPFC, rdACC)	4 42 24	0.031	62	3.63
	Instructed Group, Negative	No clusters survive FWE-correction.				
Feedback-driven prediction error	*Main effect across groups, Positive*	L Superior Orbital Gyrus	-20 26 -14	0.002	99	5.73
		L Mid Orbital Gyrus	-8 52 -10	0.016	70	4.12
	Main effect across groups, Negative	No clusters survive FWE-correction.				
	Uninstructed > Instructed	R Postcentral gyrus (BA 3b)	52 -12 30	0.028	63	4.23
	Instructed > Uninstructed	No clusters survive FWE-correction.				
	Uninstructed Group, Positive	R Postcentral Gyrus (BA 3b)	52 -12 34	0.001	106	4.56
	Uninstructed Group, Negative	No clusters survive FWE-correction.				
	Instructed Group, Positive	No clusters survive FWE-correction.				
	Instructed Group, Negative	No clusters survive FWE-correction.				

[d]. This table presents brain regions that correlate with parameters from our feedback-driven learning models (the hybrid model) within learners only. Results are whole-brain FWE-corrected at the cluster level ($p < .05$).

Table 5, Neural correlates of feedback-driven learning in the entire sample (n = 68).[e]

Learning Parameter	Contrast	Region	xyz	Cluster p(FWE-cor)	Cluster equivk	Peak equivZ
Feedback-driven associability	*Main effect across groups, Positive*	L Mid Orbital Gyrus; L Superior Medial Gyrus; L ACC (MPFC/VMPFC)	0 60 -4	0	4014	6.41
		R Amygdala; R Hippocampus	30 -4 -26	0	2508	6.2
		R Cerbellum	12 -92 -24	0	1389	5.52
		L Amygdala; L Middle Temporal Gyrus; L IFG p. Orbitalis	-28 -6 -20	0	2811	5.49
		L Cerebellum; L Inferior Temporal Gyrus	-48 -72 -30	0	278	5.49
		R Middle Temporal Gyrus; R Superior Temporal Gyrus	52 -12 -14	0	538	5.2
		Brainstem	-10 -30 -44	0	149	4.99
		L Middle Temporal Gyrus; L Inferior Temporal Gyrus	-54 -48 -6	0	180	4.95
		L Cerebellum	-42 -64 -44	0	127	4.85
		L Cuneus (V1)	2 -98 20	0	189	4.78
		R Middle Occipital Gyrus; R Middle Temporal Gyrus; R Angular Gyrus (IPL)	50 -68 28	0	203	4.7
		L ParaHippocampal Gyrus	-8 -34 -22	0	178	4.54
		L Angular Gyrus (IPL)	-40 -58 22	0	190	4.53
		L Middle Frontal Gyrus (DLPFC)	-22 10 54	0.001	110	4.53
		R Inferior Temporal Gyrus	56 -58 -10	0	179	4.43
		L Middle Frontal Gyrus; L IFG p. Triangularis	-34 52 6	0.049	60	4.35
	Main effect across groups, Negative	No clusters survive FWE-correction.				
	Uninstructed > Instructed	No clusters survive FWE-correction.				
	Instructed > Uninstructed	L Lingual gyrus; L Calcarine gyrus	-20 -78 -4	0	850	4.77
		L Middle Occipital Gyrus; L Inferior Occipital Gyrus	-40 -86 -6	0	187	4.62
		L Middle Frontal Gyrus (DLPFC)	-50 26 36	0.018	73	4.39
		L Middle Occipital Gyrus; L Superior Occipital Gyrus	-22 -88 14	0	156	4.36
		mOFC	-4 48 -32	0.002	105	4.14
	Uninstructed Group, Positive	L Insula Lobe; L Temporal Pole	-44 10 -18	0	191	5.2
		R Mid Orbital Gyrus; L ACC (MPFC; Frontal Pole)	-4 48 -2	0	267	5.05
		R Hippocampus	28 -20 -16	0.011	79	5.05
		R Amygdala	32 -4 -26	0	207	4.98
		R Cerebellum	-4 -28 -50	0	336	4.93
		Area hOc3d [V3d]	4 -98 24	0.005	91	4.92
		R Cerebellum	34 -86 -24	0	478	4.84
		R Insula Lobe; R Temporal Pole	44 16 -24	0.001	111	4.77
		R Middle Orbital Gyrus	36 40 -12	0.036	64	4.67
		L Superior Medial Gyrus	-4 60 16	0	200	4.65
		L IFG p. Orbitalis	-38 30 -12	0	267	4.48
		L Amygdala; L Hippocampus	-28 -12 -18	0.006	89	4.38
		L Middle Temporal Gyrus	-58 -18 -20	0.01	81	4.2
	Uninstructed Group, Negative	L Lingual Gyrus; L Calcarine Gyrus	-10 -76 4	0	632	4.78
		mOFC	0 46 -32	0	197	4.72
		L superior occipital gyrus; L Middle occipital gyrus	-14 -90 22	0.042	62	4.53
		Cuneus (Bilateral)	8 -82 26	0.012	78	4.01
	Instructed Group, Positive	R Superior Medial Gyrus; L Mid Orbital Gyrus; L Superior Medial Gyrus (MPFC/Frontal pole)	8 70 12	0	2558	5.87
		R Hippocampus	36 -8 -16	0	545	5.39
		R IFG p. Orbitalis; R IFG p. Triangularis (BA45/ latOFC/ latPFC)	46 34 -14	0	350	5.23
		L Cerebellum; R Precuneus; R Lingual Gyrus	-8 -46 -2	0	847	5.02
		R Middle Temporal Gyrus; R Inferior Temporal Gyrus	64 -54 -8	0	302	4.88
		L Angular Gyrus	-34 -58 24	0	129	4.79
		R Cerebellum	18 -92 -34	0.026	68	4.73
		L Amygdala; L Temporal Pole	-28 -6 -20	0	322	4.64
		L Middle Temporal Gyrus	-54 -28 -2	0	289	4.63
		R Middle Occipital Gyrus; R Calcarine Gyrus	38 -92 0	0.001	114	4.62
		L Inferior Temporal Gyrus; L Cerebellum	-50 -60 -20	0	380	4.62
		L Occipital Cortex (V3)	-34 -90 -22	0	201	4.49
		R Lingual Gyrus; R Cerebellum	16 -60 -12	0.028	67	4.48
		L Middle Temporal Gyrus	-60 -52 -8	0.001	108	4.47
		R Cerebellum	50 -76 -26	0	133	4.44
		R Middle Occipital Gyrus; R Middle Temporal Gyrus; R Angular Gyrus	48 -68 28	0.001	117	4.44
		R Middle Temporal Gyrus; R Inferior Temporal Gyrus	54 -12 -14	0	212	4.38
		L Middle Frontal Gyrus	-22 10 56	0.045	61	4.34
		R Middle Temporal Gyrus; R Superior Temporal Gyrus	58 -40 6	0	224	4.28
		L Cerebellum	-42 -64 -44	0.018	73	4.27
		L IFG p. Orbitalis (latOFC)	-44 38 -18	0	163	4.23
		R Cerebelum VI	12 -76 -20	0.042	62	4.14
		R Middle Frontal Gyrus (DMPFC); R Superior Frontal Gyrus (DMPFC)	30 34 42	0.007	86	4.08
		L Inferior Occipital Gyrus; L Middle Occipital Gyrus	-28 -94 -8	0.005	90	4.05
		L IFG p. Orbitalis; L Temporal pole	-46 22 -4	0.028	67	3.97
		R Angular Gyrus	52 -58 26	0.011	79	3.93
	Instructed Group, Negative	No clusters survive FWE-correction.				
Feedback-driven prediction error	*Main effect across groups, positive*	L Rectal Gyrus (sgACC)	-20 28 -12	0	124	5.78
	Main effect across groups, negative	No clusters survive FWE-correction.				
	Uninstructed > Instructed	No clusters survive FWE-correction.				
	Instructed > Uninstructed	No clusters survive FWE-correction.				
	Uninstructed Group, Positive	L Hippocampus	-26 -16 -16	0.003	96	4.33
		R Caudate Nucleus (VS)	8 10 -4	0.02	70	4.19
	Uninstructed Group, Negative	No clusters survive FWE-correction.				
	Instructed Group, Positive	No clusters survive FWE-correction.				
	Instructed Group, Negative	No clusters survive FWE-correction.				

[e]. Brain regions that correlate with parameters from the feedback-driven learning model across all participants. Results are whole-brain FWE-corrected at the cluster level (p < .05).

Table 6. Instructed adaptive learning in regions of interest.[d]

	Instructed Associability		Instructed Prediction Error		Instructed Expected Value		
	Left amygdala	Right amygdala	Left Ventral Striatum	Right Ventral Striatum	Left Ventral Striatum	Right Ventral Striatum	VMPFC/ OFC
ROI-based	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	t(19)= -3.44**
Voxel-based, small-volume corrected (FWE < .05) xyz coordinates	*n.s.*	*n.s.*	*n.s.*	[8 10 -6][e]	[-6 6 6]	[8 10 -6]	[-2 28 20]; [-6 38 20]

[d] This table presents ROI-based results for learning parameters derived from the modified learning model fit to behavior in the Instructed Group. ROI-based and voxel-based analyses use the same methods described in Table 1. VMPFC/OFC = Ventromedial prefrontal cortex / orbitofrontal cortex.

** *p* < .01

[e] FWE-corrected p = .051.

Table 7, Neural correlates of instructed learning within the Instructed Group learners (n = 20).[f]

Learning Parameter	Contrast	Region	xyz	Cluster p(FWE-cor)	Cluster equivk	Peak equivZ
Instructed Expected Value	*Positive correlation*	R Pallidum	8 6 2	0	291	5.5
		Thalamus, Bilateral	4 -20 2	0	1551	5.38
		R Anterior Insula; R IFG p. Orbitalis	30 28 -2	0	376	5.05
		L Anterior Insula; L Rolandic Operculum	-56 4 4	0	599	5.03
		L Middle Cingulate Cortex	-2 -2 54	0	485	4.98
		R Cerebellum	18 -46 -20	0	153	4.95
		L Superior Temporal Gyrus; L SupraMarginal Gyrus (SII; IPL)	-58 -28 26	0	227	4.63
		R Cerebellum	16 -74 -60	0.001	124	4.47
	Negative correlation	L Superior Medial Gyrus; L Superior Frontal Gyrus	-10 66 10	0	167	4.62
		R Middle Occipital Gyrus (IPL)	40 -82 18	0	199	4.57
		L Middle Occipital Gyrus	-34 -86 6	0	326	4.45
		R Lingual Gyrus; R Fusiform Gyrus	30 -44 -8	0.002	106	4.3
		L IFG p. Triangularis; L Middle Orbital Gyrus	-44 38 -2	0	161	4.22
		L Middle Frontal Gyrus (DLPFC)	-32 18 62	0	186	4.15
		R Postcentral Gyrus (BA 3b)	62 -6 24	0.013	77	4.02
		R Middle Frontal Gyrus (DMPFC)	30 24 54	0.027	68	3.89
		L Angular Gyrus (IPL)	-50 -66 34	0.034	65	3.83
Instructed Prediction Error	*Positive correlation*	No clusters survive FWE-correction.				
	Negative correlation	R Putamen; R Anterior insula; R Middle Insula	26 6 -2	0	179	4.69
		L Putamen	-30 -2 6	0	262	4.53
		R Putamen	34 -20 2	0	109	4.52
		R Supramarginal gyrus (IPL)	68 -38 30	0	155	4.33
Instructed Associability	*Positive correlation*	R Insula Lobe; R Pallidum; R IFG p. Triangularis (BA45)	34 24 -4	0	1491	4.91
		R Superior Temporal Gyrus; R Middle Temporal Gyrus; R Angular Gyrus (IPL)	60 -46 22	0	396	4.89
		R Middle Frontal Gyrus (IPL)	34 54 4	0	673	4.74
		L Anterior Insula; L IFG p. Orbitalis; L IFG p. Triangularis (BA45)	-46 18 -4	0	267	4.65
		R Middle Temporal Gyrus	58 -36 4	0.006	75	4.58
		R Superior Frontal Gyrus; R Superior Medial Gyrus (DMPFC)	22 16 56	0	280	4.52
		R Superior Medial Gyrus	4 62 22	0.027	58	4.48
		Bilateral rACC, L Superior Medial Gyrus (DMPFC)	2 44 42	0	348	4.39
		L Putamen	-18 -4 8	0.017	63	4.35
		L Cerebellum	-18 -74 -34	0	320	4.32
		R Lingual Gyris; Cerebellar Vermis	0 -72 -6	0.017	63	4.27
		R Calcarine Gyrus; R Lingual Gyrus	18 -86 8	0.021	61	4.18
		R Cerrebellum	38 -66 -28	0.004	79	4.17
		L Middle Temporal Gyrus	-48 -22 -6	0.036	55	3.84
		L Fusiform Gyrus; L Cerebellum	-42 -56 -22	0.047	52	3.78
	Negative correlation	No clusters survive FWE-correction.				

[f]. This table presents regions that correlate with parameters from our instructed learning model (the modified hybrid model) in learners in the Instructed Group. Results are whole-brain FWE-corrected at the cluster level (p < .05).

Table 8. Neural correlates of instructed learning across entire Instructed Group (n = 30).[g]

Learning Parameter	Contrast	Region	xyz	Cluster p(FWE-cor)	Cluster equivk	Peak equivZ
Instructed Expected Value	*Positive correlation*	L Thalamus	-8 -16 6	0	2459	6.38
		L Anterior Insula; L Rolandic Operculum	-32 22 6	0	1043	6.09
		R Anterior Insula	30 26 -4	0	594	5.81
		L Middle Cingulate Cortex	-2 -2 54	0	1451	5.81
		Brainstem	0 -34 -44	0.014	84	5.14
		L SupraMarginal Gyrus; L Superior Temporal Gyrus (SII; IPL)	-56 -28 28	0	1001	5.11
		R Cerebellum	20 -56 -26	0	271	5.06
		R Cerebellum	16 -74 -60	0	257	4.87
		R ACC	8 40 22	0.013	85	4.46
		L MCC	-6 -24 46	0.001	129	4.42
		R SupraMarginal Gyrus; R Superior Temporal Gyrus (IPL)	48 -32 24	0	360	4.28
		L Middle Frontal Gyrus	-38 46 40	0.005	100	4.23
	Negative correlation	L Middle Occipital Gyrus	-34 -88 14	0	588	5.24
		L Superior Medial Gyrus; L Superior Frontal Gyrus (aPFC)	-12 66 10	0	576	5.17
		R Middle Orbital Gyrus	32 42 -10	0.022	77	4.95
		L Superior Frontal Gyrus (DMPFC)	-12 46 44	0	203	4.94
		R Middle Occipital Gyrus	42 -82 14	0	356	4.59
		L Middle Frontal Gyrus (DLPFC)	-28 22 42	0	239	4.43
		L Mid Orbital Gyrus; L Rectal Gyrus (VMPFC/mOFC)	-6 54 -10	0	357	4.41
		R Postcentral Gyrus (BA3b)	64 -6 30	0.012	86	4.3
		L IFG p. Triangularis (BA45)	-48 24 16	0	174	4.28
		L Angular Gyrus (IPL)	-50 -66 34	0	142	4.28
		L Inferior Occipital Gyrus; L Middle Occipital Gyrus	-40 -78 -8	0	184	4.24
		L IFG p. Orbitalis; L Middle Orbital Gyrus (latOFC)	-44 38 -4	0	154	4.23
		R Fusiform Gyrus	30 -66 -14	0.038	69	4.19
		R Superior Medial Gyrus; R Superior Frontal Gyrus (DMPFC)	14 38 46	0	172	4.18
		R Fusiform Gyrus; R Lingual Gyrus	22 -68 -6	0	156	4.18
		L Fusiform Gyrus	-30 -58 -2	0.001	130	3.98
		R Superior Occipital Gyrus	28 -68 30	0.009	90	3.93
Instructed Prediction Error	*Positive correlation*	No clusters survive FWE-correction.				
	Negative correlation	R Supramarginal gyrus; R Angular Gyrus	62 -40 38	0	503	5.3
		R Putamen; R Anterior Insula	34 -12 2	0	1099	5.14
		L Putamen; L IFG (p Triangularis)	-26 2 0	0	1176	4.8
		L Middle Temporal Gyrus	-52 -24 -12	0.037	58	4.78
		R IFG (p Triangularis)	38 28 10	0	214	4.39
		R Posterior Insula	46 -24 -8	0.037	58	4.34
		L Cerebellum	-22 -78 -42	0.04	57	4.23
		R IFG (p Opercularis) (DLPFC)	52 14 36	0.004	86	4.12
		L Thalamus	-8 -2 16	0	123	4.05
		R Superior medial gyrus	8 34 60	0.037	58	3.77
Instructed Associability	*Positive correlation*	R Superior Frontal Gyrus; R Middle Frontal Gyrus; R IFG p. Triangularis (BA45; aPFC; latPFC)	28 54 18	0	5922	5.68
		R Superior Temporal Gyrus; R Middle Temporal Gyrus; R SupraMarginal Gyrus (IPL)	60 -46 22	0	1568	5.39
		L Cerebellum; R Calcarine Gyrus	-38 -62 -44	0	2563	5.32
		L Fusiform Gyrus; L Lingual Gyrus; L Cerebellum	-16 -56 -10	0	169	5.31
		R Superior Frontal Gyrus; R ACC (DMPFC)	12 20 58	0	1574	5.28
		R Cerebellum	36 -70 -26	0	439	4.99
		L Thalamus; L Putamen	-22 -18 4	0	206	4.96
		L Anterior Insula; L IFG p. Orbitalis	-48 18 -4	0	468	4.83
		R MCC	2 -24 28	0	192	4.65
		Midbrain	2 -14 -10	0.013	73	4.63
		L Middle Temporal Gyrus	-46 -26 -10	0.005	86	4.55
		L SupraMarginal Gyrus; L Superior Temporal Gyrus (IPL)	-64 -42 28	0	290	4.46
		R Precuneus; R Cuneus	4 -60 32	0.036	60	4.33
		L IFG p. Orbitalis	-48 32 -12	0.001	104	4.23
		L Middle Frontal Gyrus; L IFG p. Opercularis (DLPFC)	-44 16 46	0	149	4.14
		R Superior Medial Gyrus (MPFC); R rACC	4 48 2	0.002	99	4.11
		L Middle Orbital Gyrus (latOFC)	-28 56 -8	0	127	4
	Negative correlation	No clusters survive FWE-correction.				

[g]. This table presents regions that correlate with parameters from our instructed learning model (the modified hybrid model) across all participants in the Instructed Group. Results are whole-brain FWE-corrected at the cluster level (p < .05).