Neural mechanisms underlying verification and falsification in human reasoning

Jan del Ojo Balaguer[1], Maria Ruz[2] and Christopher Summerfield[1]*

[1] Dept. Experimental Psychology, University of Oxford, Oxford, UK
[2] Mind, Brain and Behaviour Research Centre, Universidad de Granada, Granada, Spain

* to whom correspondence should be addressed:
christopher.summerfield@psy.ox.ac.uk
Dept. Experimental Psychology
University of Oxford
South Parks Road
Oxford OX1 3UD, UK

Classification : Neuroscience (Biological Sciences)
Psychological and Cognitive Sciences (Biological Sciences)
Keywords : Rule-learning ; decision-making; confirmation bias ; functional MRI

ABSTRACT

Humans can learn about the world by obtaining evidence that either verifies or falsifies a contention. Where multiple theories compete to explain data, such as during scientific reasoning, falsification allows a hypothesis to be definitively ruled out, whereas verification only provides incremental evidence for one view over another. Nevertheless, using a rule-learning task, we show that human participants learn faster from evidence that verifies (rather than falsifies) a hypothesis, even though computational simulations showed that performance is maximised by learning faster from falsification. Correspondingly, functional neuroimaging revealed greater BOLD responses to verificatory relative to falsificatory feedback in the prefrontal cortex. These behavioural and neural biases were more pronounced when the number of possible hypotheses was greater. A bias to learn faster from verification may underpin suboptimal human reasoning, contributing to flawed human scientific inference.

SIGNIFICANCE

Humans are able to form hypotheses (theories) about the principles that govern the physical world. When scrutinizing data, observations can lead a theory to be verified or falsified. Where multiple theories compete to explain data, verification may increase belief in one theory over another, but falsification should prompt a theory to be definitively eliminated. We asked human participants to perform a trial-and-error rule-learning task for which performance was maximised by learning more from feedback that falsified, rather than verified, a hypothesis. Nevertheless, behavioural and neural measures showed that humans exhibited a tendency to learn preferentially from evidence that verified a contention. A bias towards verification may contribute to biased decision-making and flawed scientific inference.

\body

INTRODUCTION

The capacity to create new knowledge from experience lies at the heart of human scientific endeavour and technological advancement. Primates can select actions conditioned on conjoint information about a stimulus and its context, a capacity that underpins flexible, rule-based learning and cognitive control[1,2]. Rule-based learning occurs when clusters of sensorimotor contingencies repeatedly co-occur in a given context[3,4]. For example, after multiple visits to London, one might learn to respect the following rule: *in London, bring an umbrella*. However, humans can also use inductive reasoning to derive abstract propositional knowledge about the world (theories), which can in turn be used to make new predictions[5-7]. For example, a tourist who has never visited Oxford can draw upon their experience with the weather in London, Bristol and Edinburgh, to derive the theory *in the UK, it usually rains*, which in turn prompts the rule *in Oxford, bring an umbrella*.

According to one philosophical tradition[8], science advances as general theories are derived from observation, and evidence for a theory mounts as new phenomena are reported that verify its predictions. For example, our belief in the Standard Model of particle physics was increased by the discovery of the Higgs Boson particle[9]. Accordingly, psychologists have studied the mechanisms by which both adults and infants acquire new knowledge through induction[10], or derived metrics for imputing cause to effect on the basis of observed contingency[11-13]. However, an alternative philosophical position is that scientific knowledge is acquired not by verification but by falsification, because a limitless number of experiments are needed to verify a theory exhaustively[14]. To give one oft-cited illustration, a definitive test of the theory *all swans are white* would require the colour of all swans past, present and future to be measured, whereas this theory could be disproved by the discovery of a single black swan. This position is enshrined in conventions for the reporting of frequentist statistics, where the null hypothesis is assumed to be true unless disproved[15].

Philosophical considerations aside, how science advances will depend on the nature of the cognitive architecture that is employed by researchers as they design experiments and scrutinize data. Here, thus, we sought to understand how humans learn to reason from observation, and how verification and falsification of hypotheses are implemented in the neural circuitry underpinning voluntary choices. It has been shown that even when conclusive information is obtained only from falsification, humans seek out verificatory evidence by preference[16,17] and may overvalue such evidence when it arises[6,18,19]. Verification bias is one of a family of confirmatory strategies that allow humans to generalise old observations to new situations, but it may also hinder scientific reasoning, prompting researchers to overlook findings that are inconvenient for a favoured theory, or engage in flawed statistical methods such as circular analysis[20]. Here, we asked human participants to perform an experience-based rule learning task, allowing us to probe how humans learned from feedback that verified or falsified a current contention. Simultaneously acquiring functional magnetic resonance imaging (fMRI) allowed us to measure behavioural and neural concomitants of learning from verification and falsification, offering insights into the neural locus of the decision biases that limit human inductive inference.

Verification biases may be particularly prevalent in situations that require multiple possible hypotheses to be entertained, because capacity limits preclude the online maintenance of information about which contentions remain valid and which have been disproved. Hand-in-hand with the capacity to reason, however, humans have evolved systems of symbolic representation and communication that allow rules to be conveyed and communicated without costly trial-and-error search over possible hypotheses. A related goal of our experiment was to understand how human reasoning changes as the information burden is alleviated by symbolic information. We thus included two conditions: in the *familiar cues* condition, participants viewed symbolic cues that offered partial information about the rule that was valid in that block, whereas in the *novel cues* condition, no such information was available. Comparing behaviour and brain activity in these conditions allowed us to assess how two landmark human cognitive advances – reasoning and symbolic communication – interact during the learning of task rules.

RESULTS

Task and design. Eighteen healthy human participants learned to classify stimuli (each composed of a pair of coloured shapes occuring left and right of fixation) as 'target' or 'nontarget' on the basis of their shape (square, circle, triangle) and/or colour (red, green, blue), responding with a button press, and receiving fully informative trial-and-error feedback after each response (Fig. 1a). Each decision rule, which remained constant over a block of 16 trials, defined targets (50% of trials) as a disjunctive combination of one feature on the left and another on the right. For example, in one block stimuli were targets if the stimulus on the left was red OR the stimulus on the right was a triangle (i.e. rule={red-left, triangle-right}). The use of a disjunctive (OR rule) ensured that an effect (target) could have multiple possible causes (i.e. potential decision rules). In the previous example, positive feedback could be received for the response 'target' to the pair {red-left, circle-right}, even though circle-right is not part of the rule. This feature of the task allowed us to mimic a fundamental aspect of scientific reasoning, whereby multiple competing hypotheses vie to explain an empirically observed phenomenon.

The disjunctive task also ensured that targets and nontargets provided asymmetric information about the decision rule. By analogy with the 'black swan' example introduced above, feedback that a pair of shapes was a nontarget allowed participants to eliminate hypotheses about the decision rule component {e.g. red-left ∉ target} (falsificatory feedback), whereas information that it was a target provided only incremental evidence for {red-left ∈ target} but precluded definitive inclusion or exclusion of candidate decision rules (verificatory feedback). This design thus allowed us to pursue our main question of interest: how participants learn differently from verification (i.e. evidence indicating that a stimulus was a target) and falsification (i.e. evidence indicating that it was a nontarget), irrespective of whether feedback was positive (correct trial) or negative (error).

In the *familiar cues* condition, symbolic cues, whose meaning had been learned in a previous training session, disclosed the *class* of rule that applied in that block. The rule 'class' referred to the dimension (e.g. {colour-left, shape-right}) but not the feature (e.g. {red-left, triangle-right}) that was relevant for decisions. There were four rule classes: {colour-left, colour-right}, {colour-left, shape-right}, {shape-left, colour-right}, and {shape-left, shape-right}. In the *novel cues* condition, a distinct set of cues were paired randomly with blocks, so that they offered no information about the relevant rule. An illustration of the task is provided in Fig. 1a, and more detailed information is available in the Methods section.

Computational model. Our first concern was to understand the computational mechanisms by which humans performed the task. We began with the framework provided by reinforcement learning, assuming that participants learned associations between each state (potential cause) and a target/non-target response. In total, there were 12 potential causes: 4 side-dimension pairs (colour-right, colour-left, shape-right and shape-left) x 3 possible features (red, green, blue or circle, square, triangle), although the information provided by familiar cues reduced this space by a factor of 9. We denote these associative values $H_{i,j}$ where i and j are the indices over all the possible side-dimensions and features respectively. Model choices were made according to whether there was criterial evidence that either the left or right stimulus was a target[21].

In standard reinforcement learning models, data are fit with a single learning rate α that controls the rate at which beliefs change on the basis of new information. Here, we employed two free parameters: α_M controlled the *absolute* rate at which beliefs changed following a target or a nontarget, whereas α_R controlled the *relative* rate of belief updating. Following each trial, our model updated the associative values H for each feature present in the stimulus pair:

$$dH = 2 \cdot \alpha_M \cdot \alpha_R \cdot (+1 - H) \quad \text{if target} \tag{1.1}$$

$$dH = 2 \cdot \alpha_M \cdot (1 - \alpha_R) \cdot (-1 - H) \quad \text{if nontarget} \tag{1.2}$$

$$H = H + dH \tag{2}$$

Thus, α_R controlled whether the model learned from verification ($\alpha_R = 1$), falsification ($\alpha_R = 0$) or a mixture of the two strategies ($1 > \alpha_R > 0$). Model choices were made according to whether these values favoured a target or non-target response (see methods).

Model performance. We began by calculating how the model performed under different parameterisations of α_M and α_R, using the sequences of stimulation and feedback viewed by human participants as its input (Fig. 2a). Simulated model performance peaked close to $\alpha_R = 0$ (maximal learning from falsification), for blocks with both novel and familiar cues (Fig. 2a). Intuitively, this follows from the disjunctive rule, which ensures that falsificatory evidence is more informative. Consider a stimulus composed of a red element on the left and a blue element on the right. Feedback that the stimulus is a 'nontarget' allows the candidate rules {red-left} and {blue-right} to both be eliminated. Feedback that the stimulus is a 'target' implies that one component of the rule is either {red-left} or {blue-right}, but does not indicate which is correct. The best learning strategy is thus to weight falsificatory more than verificatory evidence ($\alpha_R \approx 0$). Model performance was relatively stable with absolute learning rate $\alpha_M > 0.2$.

Model fitting to human choice and accuracy data. Average human accuracy (% correct responses) and choices (% target responses) across the 16 trials that constituted each block are shown in Fig. 1b (dots). Accuracy increased across the block overall ($F_{(7,117)} = 29.9$, $p < 0.001$), but did so faster in blocks with familiar cues (trial x cue interaction, $F_{(7,118)} = 2.90$, $p < 0.008$). Participants began with a bias to respond 'target' that abated across the block ($F_{(7,125)} = 46.0$, $p < 0.001$) in roughly equal measure for the two conditions ($F < 1$). Next, we adjusted the model parameters to fit one half of the human choice data (even blocks), using the resulting parameters to predict performance for the remaining half (odd blocks), separately for familiar and novel conditions. The choice and accuracy of the best-fitting model parameterisation (lines) is rendered onto equivalent human data in Fig. 1b, revealing good fits across the block in each condition.

Critically, however, the parameters that allowed these fits to human performance to be obtained took on values that diverged significantly from those predicting maximal model performance (Fig. 2b). Specifically, mean values of α_R were 0.30±0.22 and 0.45±0.25 for blocks with familiar and novel cues respectively, diverging reliably from the respective parameters (α_R = 0.08 [familiar cues] and α_R = 0.11 [novel cues]) that yielded maximal performance under this model in both cases (both $t_{(17)} > 4.2$, $p < 0.001$). In other words, participants eschewed the normative strategy that would yield maximal performance, i.e. to eliminate falsified hypotheses, instead learning more from verification. Moreover, we observed a negative correlation between participants' accuracy and their best-fitting values of α_R ($r = -0.70$ and $p < 0.002$ in the familiar condition, $r = -0.31$ and $p < 0.21$ in the novel condition), indicating that those participants who learned more from falsification performed better overall (Fig. 2c). Interestingly, this deviance from optimality was accentuated in novel blocks where the space of hypothesis was bigger. This interaction was also significant ($F_{(1,17)} = 2.308$, $p = 0.147$).

Alternative models. For completeness, we also compared human performance to other candidate models. We found that an otherwise equivalent model with independent learning rates for correct and incorrect trials (CI model) accounted less well for human behavioural data. Secondly, we constructed a hierarchical Bayesian (HB) model that estimated the conditional probability that each stimulus was a target. This optimal model outperformed participants by a wide margin, and provided poorer fits to human data even when the number of parameters was taken into account. Details of model comparison are provided in Table S3.

Functional neuroimaging. Together, these behavioural and modelling analyses suggest that during inductive reasoning, humans pursue a suboptimal strategy in which learning is over-reliant on verification rather than falsification, and that this tendency is particularly pronounced in the novel cues condition, when the number of possible hypotheses about the rule exceeds the likely capacity of online maintenance processes. Next, thus, we sought to characterise the neural mechanisms which might give rise to this bias, by examining blood-oxygen (BOLD) data from functional magnetic resonance imaging (fMRI) whilst participants performed the task.

Model-based fMRI analysis. We began by using the imaging data to validate our computational model. We estimated the expected value (EV) for each trial, which reflects the probability of being correct (given the best-fitting parameterisation) and regressed this quantity against the BOLD response at each voxel in the brain (the *FV model*; Fig. 3a and methods). This analysis identified voxels in the medial orbitofrontal cortex, (or ventromedial prefrontal cortex; VMPFC) (peak: -6, 52, -2, $t_{(17)} = 6.19$, $p < 1 \times 10^{-5}$), a brain region where BOLD signals[22] and single-unit activity usually correlate with expected value[23]. By contrast, no reliable correlations with VMPFC BOLD were observed when the expected value was calculated under the performance-maximising parameterisation of the model (not shown). Furthermore, only modest ventromedial activations were observed when expected value was computed from the rival models with differential learning from correct and incorrect trials (*CI model*; Fig. 3a, centre panel) or the hierarchical Bayesian model (*HB model*; right panel), confirming the behavioural advantage for the FV model, i.e. that with differential learning from falsification and verification.

Conventional fMRI analyses. Next, we examined how VMPFC activity varied as a function of the stimulus (target vs. nontarget), and cue (novel vs. familiar). To provide a neural index of the learning occurring following feedback, we contrasted the BOLD signals elicited by correct and error feedback on these trials, using a finite impulse response (FIR) filter to estimate estimate the haemodynamic response over time (Fig. 3b). Although there was a strong effect of feedback in the VMPFC ($p < 0.0001$), the divergence between correct and error signals occurred in roughly equal measure for targets in novel and familiar cues conditions (time x cues interaction, $p = 0.38$) and for nontargets in novel and familiar cues conditions ($p = 0.33$). However, a reliable stimulus x time interaction showed that overall, there was more robust neural correlates of learning in the VMPFC from targets (i.e. verificatory feedback) than from nontargets (falsificatory feedback), irrespective of the cues (stimulus x time interaction, $F_{(5,85)} = 5.2$, $p < 0.02$; Fig. 3b left panels). When we explored other brain regions that responded robustly to correct > error trials, such as the putamen (Fig. 3b, right panel and Fig. S2), we observed no differences between the signals elicited by target/nontarget stimuli or familiar/novel cues (all p-values > 0.4).

The BOLD responses observed in the VMPFC thus provide neural evidence that participants learned faster from verification than falsification irrespective of whether the cues were novel or familiar. However, behavioural data above indicate that learning from falsification is particularly weak when the cues are novel and the number of possible hypotheses is large. To pinpoint the neural locus of this bias, we identified voxels that responded to the three-way interactions between stimulus, cue and feedback. Significant three-way interactions among were observed prominently in dorsal stream cortical areas implicated in online maintenance of task-relevant information and model-based learning (Fig. 4a), including caudal portions of the dorsolateral prefrontal cortex (DLPFC; right peak: 38, 8, 58; $t_{(17)} = 10.03$, $p < 1 \times 10^{-8}$), the rostrolateral prefrontal cortex (RLPFC; left peak: -38, 60, 2; $t_{(17)} = 5.97$, $p < 1 \times 10^{-5}$; right peak: 34, 60, 2; $t_{(17)} = 5.92$, $p < 1 \times 10^{-5}$), and the inferior parietal lobule (IPL) (right peak: 42, -44, 42; $t_{(17)} = 7.61$, $p < 1 \times 10^{-6}$; left peak: -38, -44, 38; $t_{(17)} = 6.3$, $p < 1 \times 10^{-5}$). All activations reported here survive whole-brain correction for multiple comparisons using the voxelwise false discovery rate approach[24]. A full description of the regions activated is reported in tables S1 and S2.

To understand better how these regions were responding during the task, we again re-estimated brain activity using a finite impulse response (FIR) filter, allowing us to plot peristimulus BOLD signals for each condition of the 2 x 2 x 2 design (Fig. 4b). In order to ensure independence from previous analyses, we estimated BOLD signals for each participant from a region defined from the remainder of the cohort (see methods). In each of the regions, for the familiar cues condition BOLD signals initially encoded the choice made by participants (choose target > choose nontarget) and then later encoded the feedback received (error > correct). These observations were qualified by reliable stimulus x feedback interactions (i.e. main effect of choice, target vs. nontarget) in an early time window at ~6-8s post-stimulus (IPL: $F_{(1,17)} = 26.7$, $p < 0.0001$; RLPFC: $F_{(1,17)} = 9.3$, $p < 0.008$), accompanied by a nonsignificant effect of feedback (all p-values > 0.1), alongside a significant main effect of feedback in a later window some ~8-10s post-feedback (IPL: $F_{(1,17)} = 27.8$, $p < 1 \times 10^{-4}$); DLPFC: $F_{(1,17)} = 25.4$, $p < 1 \times 10^{-4}$; RLPFC: $F_{(1,17)} = 10.7$, $p < 0.006$) with no reliable stimulus x feedback interaction (all p-values > 0.1). Although robust BOLD responses were observed in the novel cues condition, they did not differ as a function of choice or feedback in this way (all p-values > 0.1).

Comparing error and correct responses. According to reinforcement learning models, the relative brain response evoked by positive and negative feedback provides an index the effectiveness of learning in a given condition. The analysis above allowed us to explicitly compare the BOLD response for error and correct trials for each combination of stimulus (target, nontarget) and cue (familiar, novel) (Fig. 4c). In the DLPFC and IPL, for *targets* the error–correct signal diverged strongly from zero, but did not differ according to whether the cues were familiar (full lines) or novel (dashed lines) (time x cues interaction for targets: DLPFC, p = 0.41; IPL p = 0.08), although this effect did reach significance in the RLPFC ($F_{(5,83)}$ = 2.8, p < 0.03). However, for *non-targets*, the error-correct signal diverged strongly from zero for familiar but not for novel cues (time x cues interaction for nontargets: DLPFC, $F_{(5,81)}$ = 7.3, p < 0.0001; IPL, $F_{(4,72)}$ = 4.3, p < 0.003; RLPFC, $F_{(4,74)}$ = 5.9, p < 0.0001). Of note, maximal differences were observed in the later of the two time-windows, consistent with an effect driven by the feedback. In other words, for each of these regions, neural signals accompanying learning from targets were equally strong for familiar and novel cues, but neural signals indexing learning from nontargets were stronger for familiar than novel cues. This is consistent with the finding from behaviour that learning from nontargets is specifically dampened when the cues are novel and capacity is stretched to the limit.

DISCUSSION

Humans can generalise rule-based knowledge to make inferences in novel situations. For example, knowledge of the features that denote membership of a given category allows previously unseen exemplars to be accurately classified ('feature generalisation'). This allows high-level categories and concepts to be formed ('that is a dog') and contributes to linguistic development during infancy[5]. However, even in maturity humans are subject to stereotypical inductive biases, betraying a tendency to learn preferentially from information that confirms, rather than disconfirms, a currently-entertained hypothesis[17]. Failures of human induction can stymie scientific reasoning, provoking researchers to discount evidence that contradicts an established or currently-favoured hypothesis. Excessively reliance on verification in reasoning may increase the volatility of scientific progress, as evidence accretes strongly around a single theory until it is dramatically swept away in a 'paradigm shift'[25]. In other domains, confirmatory biases may be yet more pernicious, leading for example to stereotyping and prejudice on the basis of race, gender, or social background[17].

Here, we explored the neural and computational basis for a bias to learn from verification in inductive reasoning, creating a trial-and-error rule-learning task in which falsification allowed definitive inferences about a rule, whereas verification only allowed beliefs to be updated incrementally. Our task simulates the circumstances that scientists face when interpreting data, whereby the observation of an empirical effect often provides evidence for multiple competing hypotheses, whereas the falsificatory evidence can allow a theory to be definitively ruled out. We modelled the task with a learning model in which evidence that each stimulus produced an effect was updated following trial-and-error feedback. Many theories compete to describe how humans impute cause to effect, including models that rely on associative mechanisms[26] and those that assume probabilistic inference over the structure of the world[11,27]. Our experiment was not designed to speak directly to this controversy, but rather to investigate how humans weight verificatory and falsificatory evidence when reasoning about causes.

Humans performed suboptimally on the task, exhibiting a bias towards learning from verification, even though model simulations showed that falsificatory learning would have led to improved performance. Model fits demonstrated that suboptimal performance was driven by steeper learning from feedback that verified (rather than falsified) a contention. Accordingly, when we measured brain activity in the VMPFC, a brain region characterised by its robust response to informative feedback and reinforcement, the BOLD signal evoked by errors and correct trials diverged more sharply for targets (providing verification) than for nontargets (permitting falsification).

Behaviourally, the verification bias was less pronounced when an advance cue provided partial information about the rule (familiar cues condition), reducing the number of competing alternative hypotheses by a factor of two (relative to the novel cues condition). Neurally, this interaction between the target and the cue was reflected in BOLD activity recorded from dorsal stream structures previously implicated in learning and acting upon theories about the world (‘model-based’ rather than ‘model-free’ learning)[28-31]. In these regions, neural signatures of learning were particularly dampened following falsificatory feedback in the novel cues condition. In our model, efficient performance required the parallel maintenance and updating of multiple competing hypotheses about the rule, a process that is likely to draw heavily upon control structures in the parietal and prefrontal cortices. The reduction in learning-related activity in these regions following nontargets may reflect a failure of this updating process, consistent with the PFC as a locus for prediction errors on causal learning tasks[32,33]. These findings are consistent with previous brain imaging work suggesting that the prefrontal cortex maintains task rules during inductive tasks[34-36] and that patients with (left) prefrontal damage are impaired in generating abstract rules[37].

Our simulations show that on our task, learning from verification is suboptimal. Why might humans use such a strategy? A key challenge faced by any intelligent system is that the space of possible occurrences in the world is virtually unbounded, a state of affairs that places unrealistic demands on the cognitive apparatus we use to make inferences about the local environment. In philosophy, this is known as the ‘frame problem’[38]. Hypothesis-testing may have evolved to act as a tractable inference strategy, allowing inferences to be updated about the most probable occurrences in the world, avoiding costly inference about a series of impossible or improbable events, and contributing to optimal data harvesting[39]. In our experiment, we saw that symbolic cues reduced the bias to learn excessively from verification. The evolution of a system of symbolic representation, that permits hypotheses to be represented and communicated via language, is likely to have enhanced the tractability of inference by allowing humans to focus on a few probable theories without costly trial-and-error learning.

METHODS

Subjects. 18 healthy participants (6 female, 12 male; age 20-34, mean 25.0 years) were recruited into the study in accordance with local ethical guidelines. No participants reported a history of psychiatric or neurological illness, and all had normal or corrected-to-normal vision. They were paid £35 for participation in both a practice and a scanner session on two separate days.

Stimuli and task. Participants performed a rule-learning task that required pairs of shapes to be classified as “target” or “non-target” according to an unknown rule. Each block of began with the presentation of one of 8 abstract symbolic cues (Greek letters) for 3s. After a period of 2-5s (jittered), a train of 16 pairs of stimuli appeared on the left or right of the screen at approximately 3° eccentricity. Each stimulus pair remained on the screen for 3s. Each member of the pair could be a square, circle or triangle coloured red, green or blue. Participants pressed a key (training task) or response button (scanner task) at any point during the 3s presentation period to indicate whether the stimulus was a target or nontarget. Stimulus-response contingencies were fully counterbalanced across participants. Responses were followed by fully informative auditory feedback consisting of a pairs of tones with ascending (correct) or descending (incorrect) pitch (400Hz; 800z) that lasted 200ms in total. An interval of 2-6s was interposed between stimuli, during which the screen was blank. Participants completed a total of 48 blocks during training, and then a further 48 blocks for the scanner session, which occurred on a subsequent day. In the scanner, the experiment was divided into 4 runs of 12 blocks each, buttressed by lead-in and lead-out durations of 10s. Each block lasted ~17 minutes, bringing total scanning time to just over an hour.

Rules. Rules were disjunctive – for example, in one block the rule might be "if the shape on the left is red, OR the shape on the right is blue, the stimulus pair is a target". We denote this {red-left, blue-right}. We divided rules into 4 classes according to the feature that was relevant on each side (left-right}: colour-colour, colour-shape, shape-colour and shape-shape. The same feature was never selected in both cases, for example {red-left, red-right} was disallowed, leaving 6 possible rules in each class – i.e. 24 rules total. Each rule was thus repeated twice during practice and twice in the main experiment, leading to 48 blocks. Pairs of shapes were selected pseudorandomly in each block so that 8 trials were targets and 8 were nontargets, but no combinations of coloured shapes were repeated.

Cues. Four symbolic cues were randomly assigned to the four rule classes for each participant. Half of the total 48 blocks were designated 'novel cues' blocks, and the remainder were 'familiar cues' blocks. In familiar cues blocks, the symbolic cue faithfully indicated the rule class that was relevant (but not the precise rule). Participants were fully briefed as to the meaning of the cues at the beginning of the practice session, and the cues remained unchanged during the later scanner session. For the novel cues blocks, one of the remaining four cues was chosen pseudorandomly at the start of the block (irrespective of the rule class) with the only constraint that each cue was selected an equal number of times.

Behavioural analyses. We analysed accuracy and choice (target vs. nontarget) with ANOVAs testing the influence of cue condition and trial number. We also verified the influence of the past history of verificatory and falsificatory feedback on behaviour. For each feature (e.g. red on the left), we estimated the probability to respond target (as a proportion of target responses) as a function of the dimension (relevant, irrelevant), the cue (familiar, novel) and the number of times it previously appeared associated with a target (one, two, three, or more). Probabilities shown in Fig. 2d were averaged across features and sides (left and right). All ANOVAs were carried out with Greenhouse-Geisser correction for sphericity (reporting adjusted d.f. rounded to the nearest integer) using an alpha of $p < 0.05$.

Computational modelling. Our model updated beliefs about a space of possible hypotheses concerning the rule on each side of the screen, using a delta rule (equations described in the main text). In total, there were 12 hypothesis computed in parallel, defined by the dimension/side combination (colour-left, colour-right, shape-left, shape-right} and the feature (red/green/blue or shape/circle/triangle). Only the values corresponding to the features currently shown on the screen are updated. This space is reduced to 6 values in the familiar case (where only the relevant dimensions are taken into account).

On every trial, the model chose target if $TV \geq 0$, and nontarget otherwise, where:

$$TV = H_{MAX} + \tau \cdot (H_{MIN} - H_{MAX}) \quad (3)$$

where H_{MAX} (H_{MIN} respectively) is the maximum (minimum) associative value of features in the current trial, and τ was an additional free parameter that controlled whether decisions were mainly based on the most, or least, diagnostic feature information. We do not focus on analyses of τ because of space constraints; more information is given in the supplementary materials.

Fittings of the model were done using an exhaustive search through a 21x21x21 grid with uniformly distributed values over between 0 and 1 with a step of 0.05 for parameters α_M, α_R and τ. The best-fitting model was deemed to be that which minimized the MAE on choice and performance:

$$\varepsilon = 0.5 * (MAE_{CHOICE} + MAE_{PERFORMANCE}) \quad (4)$$

where MAE_{choice} and $MAE_{performance}$ are the Mean Absolute Error scores on choice and performance respectively. The model was also fit to half of the data (even blocks) and used to predict the rest (odd blocks). This allowed us to avoid over-fitting (see Table S3).

fMRI Data Acquisition and Preprocessing. Magnetic resonance images were acquired with a 3T Siemens VERIO scanner with a 32-channel head coil using a standard echo-planar imaging sequence. Whole-head T_2*-weighted echo-planar images were continuously acquired with a repetition time of 2 s, echo time of 30 ms. We acquired 510 volumes per block, plus 3 dummy scans discarded before the analyses. Each volume included 64 × 64 × 36 voxels of 3 × 3 × 3 mm. A high-resolution T_1-weighted structural image was also obtained (voxel size = 1 × 1 × 1 mm). For standard preprocessing and univariate statistical analyses, we used SPM8 (Wellcome Department of Cognitive Neurology, London, United Kingdom). All other analyses were done with custom scripts for Matlab (Mathworks, Natick, MA, United States of America). We also used xjview (http://www.alivelearn.net/xjview) to visualize the data and to construct mask and conjunction images. For each participant, we first realigned all functional images, then we co-registered (rigid body transformation) the subject's anatomical scan to the mean functional image, and then co-registered the participant's data to the Montreal Neurological Institute (MNI) template brain. We then normalized each subject's data to the template brain space, using segmented probabilistic maps for grey matter, white matter, and cerebro-spinal fluid. Functional images were resampled (4 × 4 × 4 mm voxels) and spatially smoothed (8-mm full-width half-maximum (FWHM) Gaussian kernel).

fMRI analysis. Our univariate analyses used a generalized linear model (GLM) approach. A 128-s temporal high-pass filter was applied to remove low-frequency scanner artifacts. Temporal autocorrelation in the time series data was estimated using restricted maximum-likelihood estimates of variance components using a first-order autoregressive model (AR-1), and the resulting non-sphericity was used to form maximum-likelihood estimates of the activations. Our GLM included regressors coding for onsets and durations of stimuli or events, which were then convolved with the canonical hemodynamic response function (HRF) and regressed against the observed fMRI data. Experimental blocks were modelled using separate regressors, and constant terms for each block were included. Additionally, motion parameters were included as nuisance variables.

For the model-based analyses, we constructed a design matrix with 3 regressors: the expected value (i.e. |TV|) predicted by the best-fitting parameterisation of the VF, CI and HB models. For conventioanl analyses, we included a design matrix with regressors aligned to stimulus onset, encoding the main effect of stimulus (target vs. nontarget), cues (novel vs. familiar) and feedback (correct vs. error), alongside their two- and three-way interactions. We report voxels that responded to these regressors at thresholds that were corrected for multiple comparisons, using a false discovery rate of $p<0.05$. Voxelwise statistics were rendered onto the MNI template brain using xjview (http://www.alivelearn.net/xjview8/). Subsequently, to interrogate the complex interactions among factors, we extracted raw data from cluster-corrected regions of interest (ROIs) and resubmitted them to a new finite impulse response (FIR) regression analysis that estimated the parameter estimates associated with each of 16 peri-stimulus time bins. To avoid selection bias ('double-dipping'), data for each participant were extracted on the basis of an ROI defined for the remaining 17 members of the cohort. Haemodynamic response data were temporally upsampled by a factor of 10 for plotting. Statistics are reported for an early time window (~6-8s post-stimulus) and a late time window (~8-10s post-feedback) where the BOLD signal peaked overall.

FIGURE 1

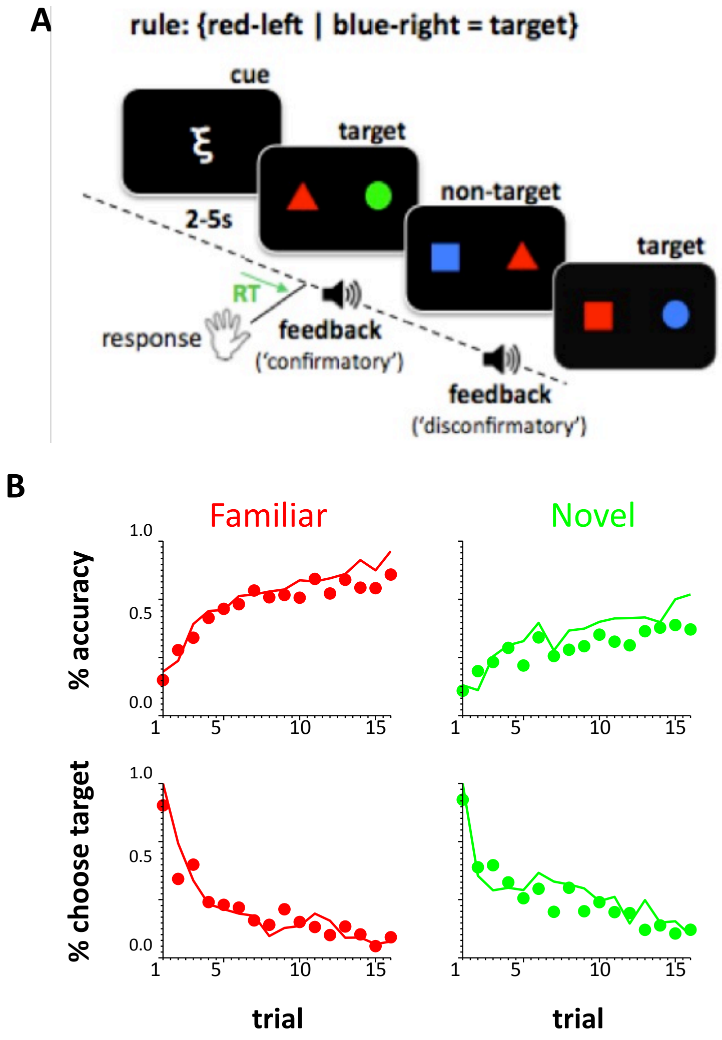

Figure 1. A. Behavioural task. On each trial participants view a pair of coloured shapes on the left and right of the screen. They responded 'target' or 'nontarget' and were provided with positive or negative auditory feedback. Each block of 16 trials was preceded by a Greek symbol which was either fully predictive of the relevant dimensions (e.g. {colour-left, colour-right}; familiar cues condition) or not at all predictive (novel cues condition). Associations between cues and dimensions were learned in a prior training session. **B.** Behavioural data from 18 human participants (dots) and predictions of the model (lines). % accuracy (top panels) and % target choice (bottom panels) over trials (1-16) are shown for familiar cues (left panels; red) and novel cues (right panels; green). Lines show the same data for the best-fitting parameterisation of the model.

FIGURE 2

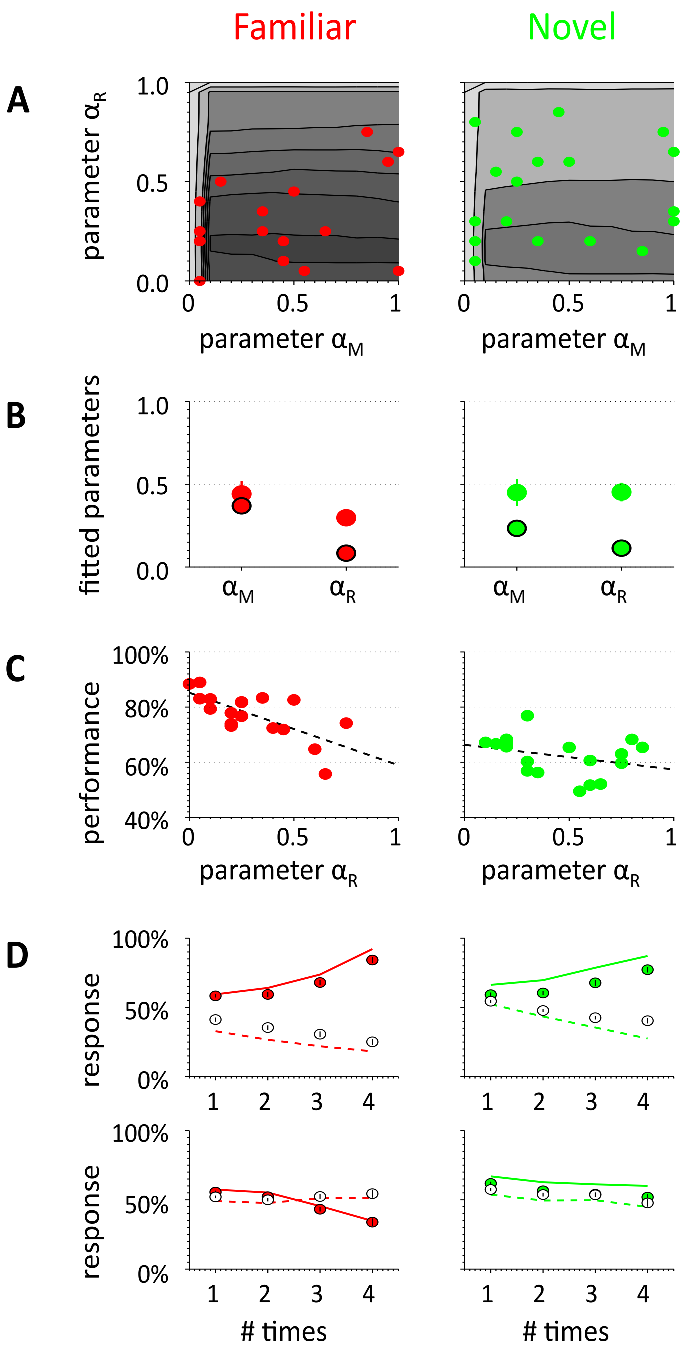

Figure 2. A. Model performance (% correct) as a function of model parameters α_M (overall learning rate) and α_R (steepness of falsificatory vs. verificatory learning) on blocks with familiar cues (left panel) and novel cues (right panel). Darker grey indicates higher performance (% correct). Red and green dots show individual humans subjects in the novel and familiar cues conditions respectively. **B.** Best-fitting (coloured) and reward-maximising (bordered in black) values of α_M and α_R in the familiar (red) and novel (green) cues condition. **C.** Correlations between α_R and performance (% correct) in the familiar (left panel) and novel (right panel) conditions for each subject. The line shows the best-fitting linear trend. **D.** % choose target responses as a function of the number of previous times the rule-relevant feature (top panels) and rule-irrelevant feature (bottom panels) was associated with confirmatory (filled circles) or disconfirmatory (open circles) feedback. Data are shown separately for familiar (left panels) and novel (right panels) cues conditions.

FIGURE 3

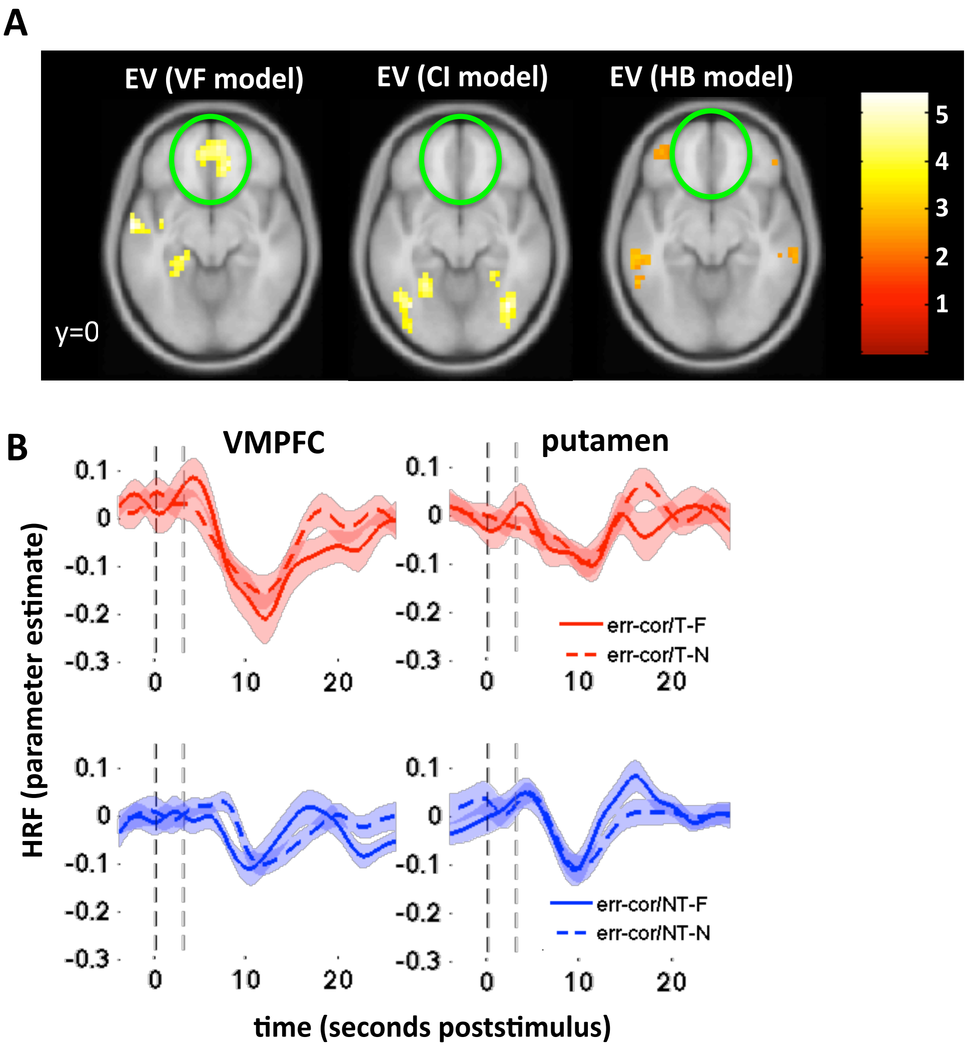

Figure 3. A. Voxels correlating with expected value predicted by the best-fitting parameterisation (left panel) and the parameterisation that maximises performance (right panels) rendered onto a transverse slice of the MNI template brain. The green ring highlights the VMPFC. **B.** Voxels correlating with the maximum hypothesis value (left panel) and minimum hypothesis value (right panel). Green rings indicate the medial and lateral orbitofrontal cortex respectively. C. Relative peri-stimulus BOLD responses on error and correct trials (error minus correct) for targets (top panels) and non-targets (bottom panels) in the VMPFC (left panel) and putamen (right panel).

FIGURE 4

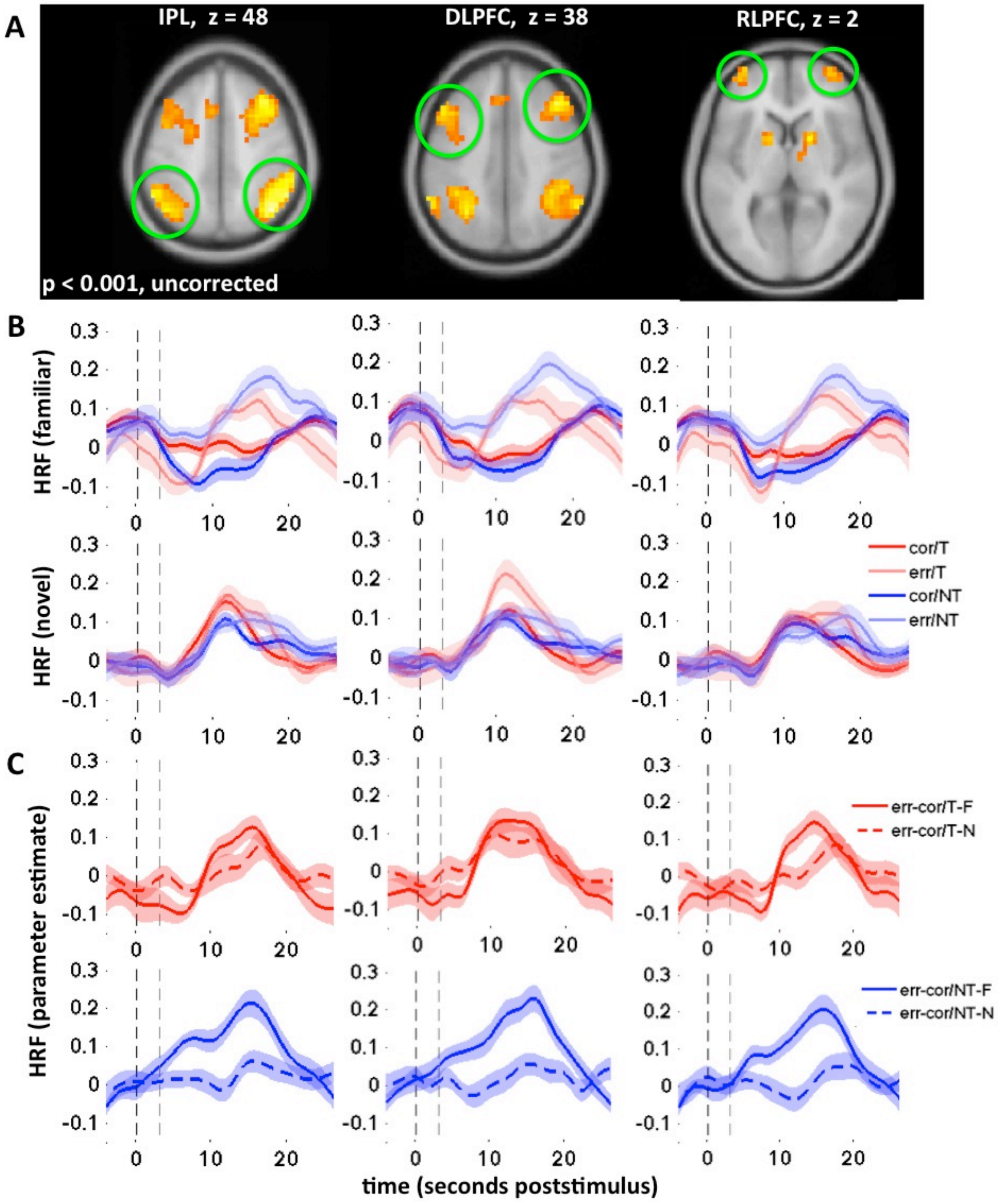

Figure 4. A. Voxels responding to the three-way interaction between stimulus (target vs. nontarget), cue (novel vs. familiar) and feedback (correct vs. error) at a threshold of p < 0.0001 (uncorrected). **B.** Haemodynamic responses estimated with an FIR filter in three brain regions: the inferior parietal lobule (IPL), dorsolateral prefronta cortex (DLPFC) and rostrolateral prefrontal cortex (RLPFC). Blue lines show nontarget trials and red lines show target trials; darker lines show correct trials and lighter lines error trials. Upper panels show the responses in familiar cues blocks, and lower panels the responses in novel cues blocks. The darker dashed line at time zero indicates stimulus onset and the subsequent lighter dashed lines shows when the auditory feedback signal occurred. **C.** Data from figure 3b are replotted as difference between error and correct trials for IPL, DLPFC and RLPFC as a function of cues (familiar [full lines] vs. novel [dashed lines]) and stimulus (target [red; upper panels] vs. nontarget [blue; lower panels]).

ACKNOWLEGDEMENTS

This work was supported by an award from the Spanish ministry of science and innovation (PSI2010-16421) to M.R. and C.S. We thank Monika Waszczuk for help with data collection, and Konstantinos Tsetsos, Hannah Tickle and José Perales for comments on the manuscript.

REFERENCES

1. Koechlin, E. & Summerfield, C. An information theoretical approach to prefrontal executive function. *Trends Cogn Sci* **11**, 229-235 (2007).
2. Miller, E.K. & Cohen, J.D. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* **24**, 167-202 (2001).
3. Miller, E.K., Freedman, D.J. & Wallis, J.D. The prefrontal cortex: categories, concepts and cognition. *Philos Trans R Soc Lond B Biol Sci* **357**, 1123-1136 (2002).
4. Toni, I., Ramnani, N., Josephs, O., Ashburner, J. & Passingham, R.E. Learning arbitrary visuomotor associations: temporal dynamic of brain activity. *Neuroimage* **14**, 1048-1057 (2001).
5. Kemp, C. & Jern, A. A taxonomy of inductive problems. *Psychon Bull Rev* (2013).
6. Bruner, J.S., Goodnow, J.J. & Austin, G.A. *A study of thinking*, (WIley, New York, 1956).
7. Heit, E. Models of inductive reasoning. in *Cambridge handbook of computational psychology* (ed. Sun, R.) 322-338 (Cambridge University Press., 2008).
8. Hume, D. *A Treatise of Human Nature (1967 edition)*, (Oxford University Press, Oxford, 1740).
9. ATLAS. Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC. *Physics Letters B* **716**, 1-29 (2012).
10. Gopnik, A. & Schulz, L. Mechanisms of theory formation in young children. *Trends Cogn Sci* **8**, 371-377 (2004).
11. Griffiths, T.L. & Tenenbaum, J.B. Structure and strength in causal induction. *Cogn Psychol* **51**, 334-384 (2005).
12. Jenkins, H.M. & Ward, W.C. Judgment of contingency between responses and outcomes. *Psychological Monographs*, 79 (1965).
13. Cheng, P.W. From covariation to causation: A causal power theory. *Psychological Review* **104**, 367–405 (1997).
14. Popper, K.R. *The Logic of Scientific Discovery*, (Basic Books, New York, 1959).
15. Fisher, R.A. *The design of experiments. 8th edition.*, (Hafner, Edinburgh., 1966).
16. Wason, P.C. Reasoning about a rule. *Quarterly Journal of Experimental Psychology* **20**, 273–281 (1968).
17. Nickerson, R.S. Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology* **2**, 175-220 (1998).
18. Doll, B.B., Hutchison, K.E. & Frank, M.J. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J Neurosci* **31**, 6188-6198 (2011).
19. Doll, B.B., Jacobs, W.J., Sanfey, A.G. & Frank, M.J. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Res* **1299**, 74-94 (2009).
20. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives in Psychological Science* **4**, 274-290 (2009).
21. Perales, J.C. & Shanks, D.R. Models of covariation-based causal judgment: a review and synthesis. *Psychon Bull Rev* **14**, 577-596 (2007).
22. Plassmann, H., O'Doherty, J. & Rangel, A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J Neurosci* **27**, 9984-9988 (2007).
23. Padoa-Schioppa, C. & Assad, J.A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223-226 (2006).

24. Genovese, C.R., Lazar, N.A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15**, 870-878 (2002).
25. Kuhn, T.S. *The Structure of Scientific Revolutions.* , (University of Chicago Press, Chicago, IL, 1962).
26. Shanks, D.R. & Dickinson, A. Associative accounts of causality judgment. in *The psychology of learning and motivation: Vol. 21. Advances in research and theory* (ed. Bower, G.H.) 229-261 (San Diego: Academic Press, 1987).
27. Holyoak, K.J. & Cheng, P.W. Causal learning and inference as a rational process: the new synthesis. *Annu Rev Psychol* **62**, 135-163 (2011).
28. Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P. & Dolan, R.J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204-1215 (2011).
29. Glascher, J., Daw, N., Dayan, P. & O'Doherty, J.P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585-595 (2010).
30. Dolan, R.J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312-325 (2013).
31. Smittenaar, P., Fitzgerald, T.H., Romei, V., Wright, N.D. & Dolan, R.J. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron* (2013).
32. Corlett, P.R.*, et al.* Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron* **44**, 877-888 (2004).
33. Turner, D.C.*, et al.* The role of the lateral frontal cortex in causal associative learning: exploring preventative and super-learning. *Cereb Cortex* **14**, 872-880 (2004).
34. Crescentini, C.*, et al.* Mechanisms of rule acquisition and rule following in inductive reasoning. *J Neurosci* **31**, 7763-7774 (2011).
35. Hampshire, A., Thompson, R., Duncan, J. & Owen, A.M. Lateral prefrontal cortex subregions make dissociable contributions during fluid reasoning. *Cereb Cortex* **21**, 1-10 (2011).
36. Goel, V. & Dolan, R.J. Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition* **93**, B109-121 (2004).
37. Reverberi, C., D'Agostini, S., Skrap, M. & Shallice, T. Generation and recognition of abstract rules in different frontal lobe subgroups. *Neuropsychologia* **43**, 1924-1937 (2005).
38. McCarthy, J. & Hayes, P.J. Some philosophical problems from the standpoint of artificial intelligence. in *Machine intelligence*, Vol. 4 (eds. Meltzer, B. & Michie, D.) (Edinburgh University Press, 1968).
39. Oaksford, M. & Chater, N. Precis of bayesian rationality: The probabilistic approach to human reasoning. *Behav Brain Sci* **32**, 69-84; discussion 85-120 (2009).

SUPPLEMENTARY MATERIALS

Supplementary Section 1 - Model policy

Our task followed a disjunctive rule (e.g., if {red-left} or {blue-right} then target). An optimal policy would thus respond target if any of the features presented in the current trial satisfies the rule. This can be mathematically translated as an evaluation over H_{MAX}, the maximum associative value corresponding to any of these features. Additionally, an optimal parameter $\alpha_R = 0$ would ensure that all H values are either -1 or 0. Thus, an optimal policy would be target if $H_{MAX} \geq 0$ and nontarget otherwise.

For other parametrisations $\alpha_R > 0$, the range of H values is between -1 and +1 and depends on the trial number. For instance, with α_R close to 1, the optimal policy described above would perform poorly because most H values would be positive (thus respond always target).

We defined a parameterised extension of the optimal policy in order to achieve better leverage, and to assess the optimality of human's policy : we evaluated to which extent humans made use of a suboptimal policy based on H_{MIN} (the minimum associative value in the current trial). The parameter τ allowed us to measure to which extent humans made use of each policy.

As expected, simulations on the reward-maximising VF model based its policy on H_{MAX} (values of τ were 0.06±0.03 and 0.08±0.04 for familiar and novel respectively). Similarly to the results we found for α_R, fittings to τ showed a significant deviance from optimality (both $t_{(17)} > 4.06$ and $p < 0.001$ for familiar and novel conditions), and an interaction ($F_{(1,17)} = 3.048$, $p < 0.099$) where this deviance was accentuated in the novel condition.

Additionally, we found a strong correlation between parameters that was mainly driven by a linear relation between α_R and τ ($r = 0.91$, $p < 10^{-6}$ in the familiar condition; $r = 0.97$, $p < 10^{-12}$ in the novel condition).

Supplementary section 2 - Hierarchical Bayesian model.

Notation

For simplicity will note P(t), P(x) and P(r) the probabilities P(T=1), P(X=x) and P(R=r) respectively, where T is the random binary variable describing if any given trial is target (T=1) or nontarget (T=0); R is a random variable associated to the block rule; and X is a variable associated with the stimulus set presented in a trial. A subscript $P(t_n)$, $P(x_n)$ will be used to refer to this probability for trial n. We will use **bold** notation $\mathbf{x_n}=(x_1...x_n)$ and $\mathbf{t_n}=(t_1...t_n)$ to refer to the history of previous trials (stimulus and target respectively). The same notation will be used to refer to conditional probabilities.

Assumptions

Our hierarchical Bayesian (HB) model is built under a few simple assumptions. Like humans, the model is given the space of possible rules and the output (target or nontarget) that each rule predicts for each pair of stimuli presented during a trial. The HB model doesn't take into account the fact that each block had 50% target trials and 50% nontarget trials (neither were humans informed of this). The HB model was simulated independently on each different block. Thus, the history of rules in previous blocks wasn't exploited to predict the current rule. Another prior was that the underlying rule r and any observation x_i were uniformly distributed across their respective spaces. The probability for the current trial of being target was thus dependent on the probability of the underlying rule and the history of trials within the block, but not the type of block (i.e., which were the relevant dimensions). In familiar blocks, rules associated with the irrelevant dimensions were discarded (see Implementation). The hierarchical Bayesian model is optimal based on the following assumptions:

- independence of block rules across blocks
- independence of the stimulus presented with the history of previous stimuli
- independence of the stimulus presented with the underlying rule
- knowledge about the space of possible rules
- knowledge about the constraint due to the block cues
- uniform distribution across rules
- uniform distribution across observations

Structure

The HB model uses two layers: the target layer and the rule layer.

The first «target» layer estimates the probability for a certain trial of being target: $P(t_n|\mathbf{x_n},\mathbf{t_{n-1}})$. This is calculated as the marginal probability across all the possible rules:

$$P(t_n|\mathbf{x_n},\mathbf{t_{n-1}}) = \Sigma_r \{ P(t_n|r,x_n)\, P(r|\mathbf{x_{n-1}},\mathbf{t_{n-1}}) \} \qquad \text{(eq. SE1)}$$

where Σ_r is the sum across all possible rules (i.e., the 6x6 combinations of features between left and right sides). Note that the target probability $P(t_n|r,x_n)$ was only dependent on the underlying rule and the current observation, while the probability of a certain rule $P(r|\mathbf{x_{n-1}},\mathbf{t_{n-1}})$ depended on the history of previous trials.

The second «rule layer» estimates, from the history of previous trials, the probability for each rule of being the underlying one: $P(r_n|\mathbf{x_{n-1}},\mathbf{t_{n-1}})$. For instance, it can be shown that

$$\begin{aligned}
&P(r|\mathbf{x_{n-1}},\mathbf{t_{n-1}})\\
&\quad = P(\mathbf{t_{n-1}}|r,\mathbf{x_{n-1}}) / \Sigma_r \{ P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}}) \}\\
&\quad = P(r)\, P(\mathbf{x_{n-1}} , \mathbf{t_{n-1}} \mid r) / P(\mathbf{x_{n-1}} , \mathbf{t_{n-1}})\\
&\quad = P(r)\, P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}}) / P(\mathbf{t_{n-1}} \mid \mathbf{x_{n-1}})\\
&\quad = P(r)\, P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}}) / \Sigma_r \{ P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}})\, P(r \mid \mathbf{x_{n-1}}) \}\\
&\quad = P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}}) / \Sigma_r \{ P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}}) \}\\
&\quad = \Pi_{i=1..n-1} \{ P(t_i \mid r , x_i) \} / \Sigma_r \{ P(\mathbf{t_{n-1}} \mid r , \mathbf{x_{n-1}}) \} \qquad \text{(eq. SE2)}
\end{aligned}$$

where $\Pi_{i=1..n-1}$ is the product operator, and under the additional assumptions that $\mathbf{x_n}$ and r are independent, and x_i is independent across different trials i.
This probability is 0 if the rule is inconsistent with the history of previous trials, and is equally distributed across all remaining (consistent) rules. For example, if four possible rules were consistent with the history of trials, the probability for each of these would be of 0.25, while 0 for any other rule.

This model is shown to be optimal based on the previous priors.

Implementation

We created an instance of this model by keeping track of all possible rules in a 6*6 binary matrix – called the *candidates matrix* **C**. This matrix allowed us to keep track of the rules that were consistent in previous trials. Each cell (i,j) in this matrix was associated with each possible rule on each side, where i,j ∈ {1...6} corresponded to the feature on the left and right sides respectively (e.g., red, green, blue, square, triangle, circle).

The *candidates matrix* **C** was used in the following way. If $\mathbf{C_{ij}}$=1, then the rule (i,j) was consistent with the history of previous trials – and was a candidate for the underlying rule. If $\mathbf{C_{ij}}$=0, then the rule (i,j) was inconsistent with the history of previous trials and could not candidate as the underlying rule for the current block.
The *candidates matrix* corresponds to the «rule layer» in our hierarchical Bayesian model.

At the beginning of the block, all cells in the candidates matrix were set to 1, meaning that all rules were possibly the underlying rule for the current block. In familiar blocks, the 27 rules corresponding to the irrelevant dimensions were set to 0, leaving only 9 possible rules left.

On each trial and for each candidate rule (i.e., for each rule r=(i,j) such that $\mathbf{C_{ij}}$=1), we calculated the conditional probability of being target: $P(t_n|r,x_n)$. This probability could be 1 (i.e., the current trial is target) or 0 (trial is nontarget) following the prediction given by each rule. From equation (SE1), the probability of a target trial, $P(t_n|\mathbf{x_n},\mathbf{t_{n-1}})$, was the average across these probabilities. This calculation corresponds to the «target layer» of the hierarchical Bayesian model.

We used a greedy policy for the model, where the response was target if $P(t_n)\geq 0.5$ and nontarget in any other case.

As can be seen in Fig S1 panel A, the model largely outperforms both the reinforcement model presented in equations (1–3) and % accuracy achieved by humans, by inferring the probability for each trial of being target in a Bayesian fashion. The number of consistent rules explaining the history of previous trials (see Fig S1, panel B) decreases from 9 and 36 candidates (familiar and novel blocks, respectively) to 1, when the model can conclude with complete certainty what is the underlying rule for that block. This model thus allows us to estimate an upper boundary on the accuracy of responses given by either human behaviour or any other model predictions.

Supplementary section 3 - Models based on performance.

Model CI

Analysis of human behaviour revealed a strong correlation between target and performance on a trial0by-trial level ($r = -0.0652$, $p < 10^{-13}$), due to the bias on choice at the beginning of the block. Thus, one possible confound in our experiment was between target and accuracy, i.e. effects on the VF model could result from learning from correct/incorrect trials rather than from target/nontarget.

To assess this problem, we defined an alternative model CI that used two different learning rates to update its belief on correct and incorrect trials respectively:

$$\begin{array}{lll} dH = 2 \cdot \alpha_M \cdot \alpha_R \cdot (+1 - H) & \text{if target and correct} & \text{(S 3.1.1)} \\ dH = 2 \cdot \alpha_M \cdot (1 - \alpha_R) \cdot (+1 - H) & \text{if target and incorrect} & \text{(S 3.1.2)} \\ dH = 2 \cdot \alpha_M \cdot \alpha_R \cdot (-1 - H) & \text{if nontarget and correct} & \text{(S 3.1.3)} \\ dH = 2 \cdot \alpha_M \cdot (1 - \alpha_R) \cdot (-1 - H) & \text{if nontarget and incorrect} & \text{(S 3.1.4)} \end{array}$$

Fitting values of α_R were 0.31±0.21 in the familiar condition and 0.44±0.27 in the novel condition, diverging significantly from optimal values (0.13 and 0.23 respectively; both $t_{(17)} > 3.23$ and $p < 0.005$). Model CI was used to assess the validity of model VF in predicting BOLD signal (see Figure 3). A significant correlation was found between α_R and performance for familiar and novel conditions (both $r < -0.59$, $p < 0.01$).

Model VFCI

As a complementary analysis, we also defined a more sofisticated model VFCI that allowed for simultaneously estimate the effect of correct/incorrect and target/nontarget trials in learning rates. This model was a natural extension of models VF and CI :

$$\begin{array}{lll} dH = 4 \cdot \alpha_M \cdot \alpha_R^{VF} \cdot \alpha_R^{CI} \cdot (+1 - H) & \text{if target and correct} & \text{(S 3.2.1)} \\ dH = 4 \cdot \alpha_M \cdot \alpha_R^{VF} \cdot (1 - \alpha_R^{CI}) \cdot (+1 - H) & \text{if target and incorrect} & \text{(S 3.2.2)} \\ dH = 4 \cdot \alpha_M \cdot (1 - \alpha_R^{VF}) \cdot \alpha_R^{CI} \cdot (-1 - H) & \text{if nontarget and correct} & \text{(S 3.2.3)} \\ dH = 4 \cdot \alpha_M \cdot (1 - \alpha_R^{VF}) \cdot (1 - \alpha_R^{CI}) \cdot (-1 - H) & \text{if nontarget and incorrect} & \text{(S 3.2.4)} \end{array}$$

We repeated the same analysis with model VFCI.
Fitting values of α_R^{VF} were 0.34±0.21 in the familiar condition and 0.49±0.27 in the novel condition, diverging significantly from optimal values (0.21±0.08 and 0.18±0.08 respectively; both $t_{(17)} > 3.01$ and $p < 0.008$). Fitting values of α_R^{CI} were 0.41±0.27 in the familiar condition (not significantly diverging from optimal values 0.33±0.07; $t_{(17)} = 1.20$, $p < 0.25$) and 0.48±0.25 in the novel condition (significantly diverging from optimal values 0.26±0.08; $t_{(17)} > 4.16$ and $p < 0.001$).
No significant correlation was found between α_R^{VF} (or α_R^{CI}) and performance for familiar and novel conditions.

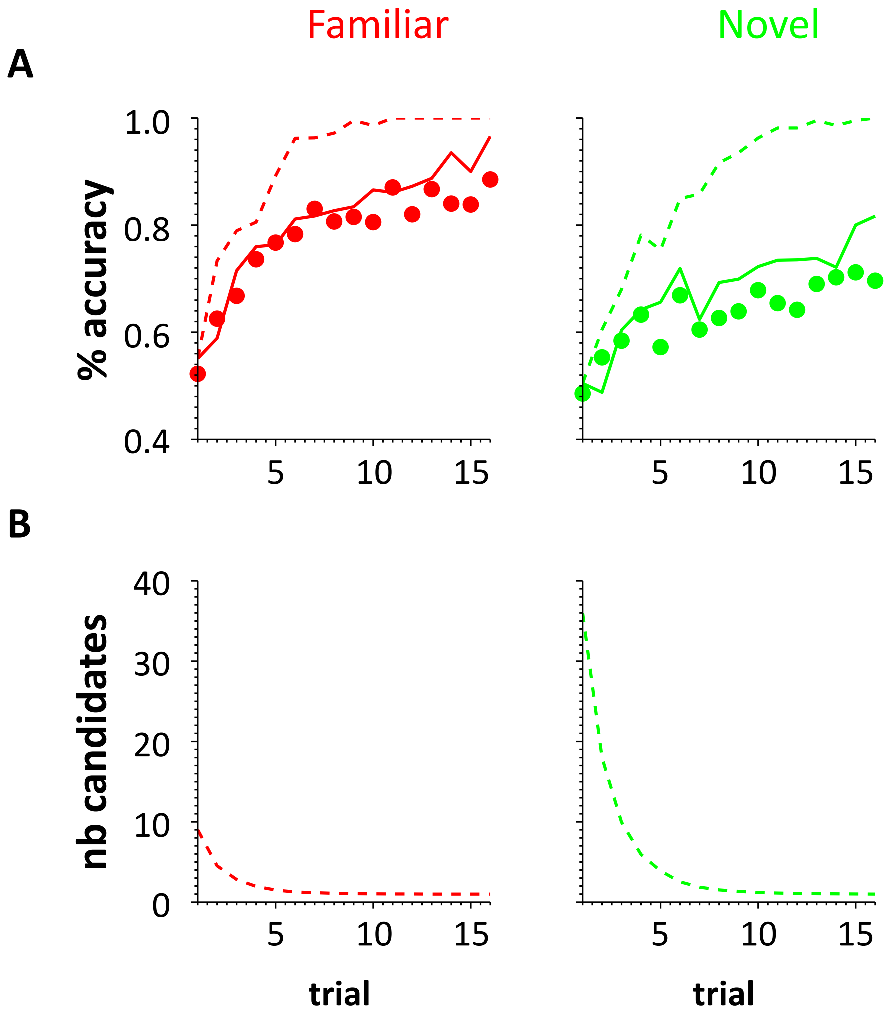

Figure S1. A. Behavioural data from 18 human participants (dots) and predictions of the reinforcement model (continued lines) and the hierarchical Bayesian model (dotted lines). % accuracy over trials (1-16) is shown for familiar cues (left panels; red) and novel cues (right panels; green). Lines show the same data for the best-peformance parametrisation of the model. **B.** Average number of candidate rules in the hierarchical bayesian model across trials, both for familiar and novel blocks (red; green). At the beginning of the block, 9 and 36 candidate rules were used in familiar and novel bocks respectively. These values decreased to 1 by the end of the block.

Figure S2. BOLD responses to the 3-way interaction between stimulus, cue and feedback.

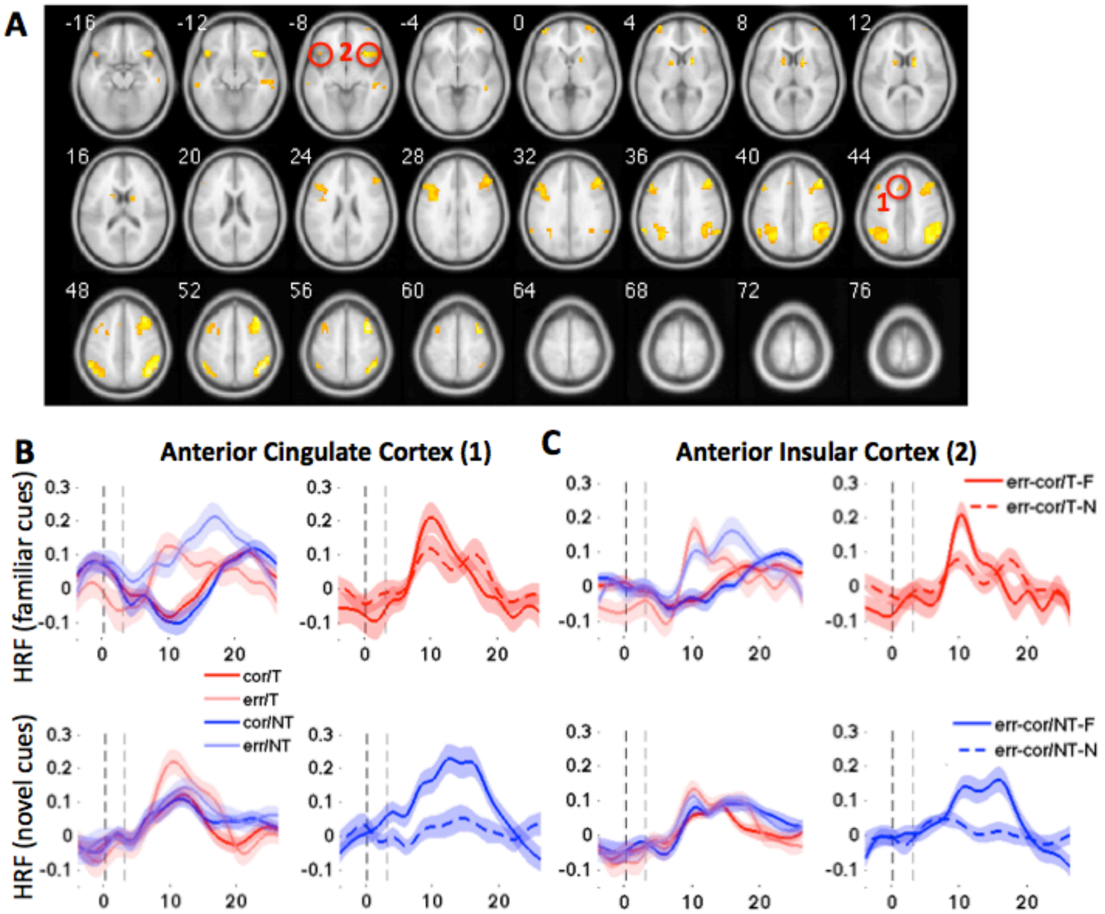

Figure S2. A. All voxels responding to the three-way interaction between stimulus, cue and feedback, rendered onto a template brain at a threshold of p < 0.001 uncorrected. All voxels shown survive false discovery rate correction at p < 0.05. Red circles and numbering indicate the locations of the dACC and AINS. **B.** Left panels: BOLD responses in the dorsal anterior cingulate cortex (ACC) to correct and error target and nontarget trials, for novel cues (bottom panel) and familiar cues (top panel) blocks. Right panels: replotting the same data as error-correct BOLD for targets (top panel) and nontargets (bottom panel). **C.** As B, but for the anterior insular cortex (AINS).

Table S1. Voxels sensitive to three-way interaction between stimulus, cue and feedback.

STATISTICS: p-values adjusted for search volume

cluster	voxel	voxel	voxel	voxel	voxel		
p(cor)	k	p(FDR)	T	Z	p(unc)	x,y,z {mm}	
0.000	395	0.000	10.03	5.66	0.000	38 8 58	right DLPFC
		0.000	8.84	5.34	0.000	38 24 54	
		0.000	7.96	5.07	0.000	46 28 42	
0.000	329	0.000	9.05	5.40	0.000	42 -64 50	right IPL
		0.000	8.89	5.36	0.000	50 -48 54	
		0.000	7.61	4.96	0.000	42 -44 42	
0.000	85	0.000	8.26	5.17	0.000	38 20 -10	
0.000	220	0.000	8.06	5.10	0.000	-10 -80 -26	
		0.000	6.60	4.59	0.000	18 -84 -34	
		0.001	5.58	4.15	0.000	14 -80 -26	
0.000	266	0.000	7.40	4.88	0.000	-30 -60 42	left IPL
		0.000	7.17	4.80	0.000	-54 -52 50	
		0.000	6.89	4.70	0.000	-42 -64 50	
0.000	63	0.000	7.19	4.81	0.000	18 8 10	right caudate
		0.006	4.24	3.45	0.000	10 -8 6	
0.000	341	0.000	6.62	4.59	0.000	-42 20 30	left DLPFC
		0.000	6.47	4.54	0.000	-42 8 30	
		0.001	6.43	4.52	0.000	-50 20 38	
0.003	35	0.000	6.55	4.57	0.000	-42 20 -10	
0.000	49	0.001	5.97	4.33	0.000	-38 60 2	left RLPFC
0.000	83	0.001	5.96	4.32	0.000	58 -24 -10	
		0.001	5.81	4.25	0.000	46 -28 -6	
		0.001	5.46	4.09	0.000	46 -36 -2	
0.000	90	0.001	5.92	4.30	0.000	34 60 2	right RLPFC
		0.001	5.57	4.15	0.000	50 36 -22	
		0.004	4.55	3.63	0.000	42 44 -10	
0.000	52	0.001	5.74	4.22	0.000	-14 8 10	left caudate
		0.003	4.87	3.80	0.000	-18 -4 22	
0.002	38	0.002	5.31	4.02	0.000	-62 -28 -6	
0.000	49	0.002	5.22	3.98	0.000	-10 20 46	
0.007	29	0.002	5.08	3.91	0.000	42 -72 -30	
		0.004	4.69	3.71	0.000	30 -64 -34	

Table S2. Voxels sensitive to the main effect of correct > error.

```
STATISTICS: p-values adjusted for search volume
============================================================
cluster  voxel   voxel   voxel  voxel   voxel
p(cor)     k     p(FDR)  T        Z     p(unc)   x,y,z {mm}
--------------------------------------------------------------------------------
0.000   189     0.022   6.79    4.66    0.000    14 -40  62
                0.025   5.54    4.13    0.000    30 -44  70
                0.031   4.54    3.63    0.000     6 -16  42
0.000   90      0.022   6.51    4.55    0.000   -34 -88  26
                0.022   6.05    4.36    0.000   -22 -92  30
                0.025   5.39    4.06    0.000   -22 -96  22
0.000   78      0.022   5.83    4.26    0.000    30  -8   2       right putamen
                0.031   4.48    3.59    0.000    38 -16  18
0.000   102     0.025   5.51    4.12    0.000    46 -80  22
                0.039   3.89    3.25    0.001    42 -68   6
0.000   51      0.025   5.23    3.98    0.000   -34   0  10
                0.033   4.33    3.51    0.000   -30 -16  10
                0.035   4.21    3.44    0.000   -34  -8  -2       left putamen
0.006   30      0.025   5.02    3.88    0.000   -14 -44  66
                0.037   4.07    3.35    0.000   -22 -48  58
0.084   15      0.025   4.94    3.84    0.000    26 -40 -22
                0.031   4.59    3.65    0.000    34 -48 -10
0.000   73      0.027   4.83    3.78    0.000   -14  48  -2       VMPFC
                0.032   4.43    3.56    0.000     6  56  14
                0.033   4.35    3.52    0.000    -6  56  18
0.005   31      0.029   4.75    3.74    0.000   -18 -44 -18
                0.032   4.39    3.54    0.000   -26 -40 -22
                0.035   4.22    3.44    0.000   -26 -36 -10
============================================================
```

Table S3. MAE scores for models.

	Fitting		**Prediction**	
	Familiar	**Novel**	**Familiar**	**Novel**
HB	0.1134 ± 0.0122	0.1813 ± 0.0113	0.1192 ± 0.0149	0.1887 ± 0.0102
VF	0.0825 ± 0.0059	0.1136 ± 0.0058	0.1128 ± 0.0080	0.1479 ± 0.0074
CI	0.0842 ± 0.0053	0.1104 ± 0.0041	0.1163 ± 0.0087	0.1534 ± 0.0072
VFCI	0.0790 ± 0.0049	0.1052 ± 0.0038	0.1155 ± 0.0079	0.1481 ± 0.0068

Model VFCI had better fittings than model VF (both $t_{(17)} > 2.7$ and $p < 0.006$) and model CI (both $t_{(17)} > 3.3$ and $p < 0.006$), but predictions were not statistically better for any model between models VF, CI and VFCI.

Compared to model HB, all three models VF, CI and VFCI achieved better fittings (all $t_{(17)} > 3.1$ and $p < 0.003$ in the familiar condition ; all $t_{(17)} > 6.6$ and $p < 0.001$ in the novel condition). Only better predictions were achieved in the novel condition (all $t_{(17)} < 0.6$ in the familiar condition ; all $t_{(17)} > 3.7$ and $p < 0.001$ in the novel condition).