Title:

Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing

Authors:

Stefano Palminteri[1,2], Germain Lefebvre[2], Emma J. Kilford, Sarah-Jayne Blakemore[1]

Affiliations:

[1]Institute of Cognitive Neuroscience, University College London, London, UK.

[2]Laboratoire de Neurosciences Cognitives, École Normale Supérieure, Paris, FR.

Corresponding author:

Stefano Palminteri (stefano.palminteri@ens.fr)

Abstract

Previous studies suggest that *factual* learning, that is learning from obtained outcomes, is biased, such that participants preferentially take into account positive, as compared to negative, prediction errors. However, whether or not the prediction error valence also affects *counterfactual* learning, that is, learning from forgone outcomes, is unknown. To address this question, we analysed the performance of two cohorts of participants on reinforcement learning tasks using a computational model that was adapted to test if prediction error valance influences learning. Concerning factual learning, we replicated previous findings of a valence-induced bias, whereby participants learned preferentially from positive, relative to negative, prediction errors. In contrast, for counterfactual learning, we found the opposite valence-induced bias: negative prediction errors were preferentially taken into account relative to positive ones. When considering valence-induced bias in the context of both factual and counterfactual learning, it appears that people tend to preferentially take into account information that confirms their current choice. By documenting these valence-induced learning biases, our findings demonstrate the presence of a confirmation bias in human reinforcement learning.

Introduction

Goal-directed behaviour is composed of two core components[1]: one component is the decision-making process that starts from representing the available options and terminates in selecting the option with the highest expected value; the second component is reinforcement learning (RL), through which outcomes are used to refine value expectations in order to improve decision-making. Human decision-making has been shown to be subject to biases (i.e. deviations from the normative prescriptions), such as the framing effect[2]. Whereas the investigation of decision-making biases has a long history in economics and psychology, learning biases have been much less systematically investigated[3]. This is surprising as most of the decisions we deal with in everyday life are experience-based and choice contexts are often recurrent, thus allowing learning to occur and influence decision-making. In addition, it is important to investigate learning biases as RL processes might play in psychiatric pathogenesis and economic maladaptive behaviour[4,5].

Standard RL algorithms learn action-outcome associations directly from obtained outcomes on a trial and error basis[6]. We refer to this direct form of learning as “factual learning”. Despite the fact that standard models, built around the notion of computational and statistical optimality, prescribe that an agent should learn equally well from positive and negative obtained outcomes, previous studies have consistently shown that humans display a significant valence-induced bias[7,8]. The bias generally goes in the direction of preferential learning from positive, compared to negative outcome prediction errors[9–12]. This asymmetry in the effects of valence on RL could represent a “low-level” counterpart of the “good news/bad news” effect observed for “high-level” real life prospects, which has been suggested to contribute to maintaining an optimism bias[13].

However, human RL cannot be reduced simply to learning from obtained outcomes. Other sources of information can be successfully integrated in order to improve performance and RL has a multi-modular structure[14]. Amongst the computational modules that have been demonstrated in humans is counterfactual learning**.** Counterfactual learning refers to the ability to learn from forgone outcomes (i.e. the outcomes of the option(s) that were not chosen)[15,16]. So far, whether or not a similar valance-induced bias also affects counterfactual remains unknown.

To address this question, we ran two experiments implicating instrumental learning and computational model-based analyses. Two cohorts of healthy volunteers performed variants of a repeated two-armed bandit task involving probabilistic outcomes[17,18] (**Figure 1A** and **1B**). We analysed the data using a modified version Rescorla-Wagner model that assumes different learning rates for positive and negative, factual and counterfactual, prediction errors (**Figure 1C**)[19,20].

The first experiment aimed to replicate previous findings of a “positive valence bias” at the level of factual learning. In this first experiment, participants were presented only with the obtained outcome (chosen outcome: R_C; **Figure 1A**)[9]. In the second experiment, in order to investigate whether or not

counterfactual learning rates are also affected by the valence of prediction errors, we used a variant of the same instrumental learning task, in which participants were also presented with the forgone outcome (unchosen outcome: R_U; **Figure 1B**). Our design allowed us to test competing hypotheses concerning the effect of valence on counterfactual learning (**Figure 2A**). A first hypothesis was that, as opposed to factual learning, counterfactual learning would be unbiased. The second hypothesis was that factual and counterfactual learning would present the same valence-induced bias, such that positive counterfactual prediction errors were more likely to be taken into account than negative counterfactual prediction errors. In this scenario the factual and counterfactual learning biases would be consequences of a more general "positive valence" bias, in which positive prediction errors have a greater impact on learning, regardless of whether the option was chosen or not. Finally, the third hypothesis was that valence would affect factual and counterfactual learning in opposing directions, such that negative unchosen prediction errors are more likely to be taken into account than positive unchosen prediction errors. In this last scenario the factual and counterfactual learning biases would be consequences of a more general "confirmation bias", in which outcomes that support the current choice, are preferentially taken into account.

Learning rate analysis, as a function of both outcome valence (positive and negative) and outcome type (chosen and unchosen), was consistent with this last hypothesis. Behavioural analysis showed that the factual and counterfactual learning biases might be maladaptive, especially in situations involving changes in reward contingencies.

Results

Behavioural task and computational models

To investigate both factual and counterfactual reinforcement learning biases we designed an instrumental task based on a previous design, in which we showed a significant optimistic bias in factual learning[9]. The two experiments were different in that the task used in Experiment 1 (N=20) involved participants being shown only the outcome of their chosen option (**Figure 1A**), whereas in Experiment 2 (N=20) the outcome of the unchosen option was also displayed (**Figure 1B**). To test our hypotheses concerning valence-induced learning biases (**Figure 2A**) we fitted the data with a Rescorla-Wagner model assuming different learning rates for positive and negative outcomes, which respectively elicit positive and negative prediction errors (**Figure 1C**). The algorithm used to explain Experiment 1 data involved two learning rates for obtained outcomes ($\alpha_c^+$and α_c^- for positive and negative prediction errors of the obtained outcomes, respectively); in addition to the obtained outcomes learning rates, the algorithm used to explain Experiment 2 data also involved two learning rates for forgone outcomes (α_u^+ and α_u^- for positive and negative prediction errors of the forgone outcomes, respectively).

Learning rate analysis

Replicating previous findings, in Experiment 1 we found that the positive factual learning rate (α_{C+}) was significantly higher than the negative one (α_{C-}; T(19)=2.4; P=0.03) (**Figure 2B**, left). Regarding Experiment 2, we submitted learning rates to a repeated-measure ANOVA with prediction error valence (positive or negative) and prediction error type (factual or counterfactual) as within-subjects factors. Falsifying the "positive valence bias" hypothesis the ANOVA revealed no main effect of prediction error valence (F(1,19)=0.5; P>0.4). We also did not find any effect of outcome type, indicating that, on average, factual and counterfactual learning were similar (F(1,19)=0.2; P>0.6). Consistent with the "confirmation bias" hypothesis we found a significant interaction between valence and type (F(1,19)=119.2; P=1.3e-9). Post-hoc tests indicated that the interaction was driven by effects of valence on both factual ($\alpha_{C+}>\alpha_{C-}$; T(19)=6.2; P=6.9e-6) and counterfactual learning rates ($\alpha_{U+}>\alpha_{U-}$; T(19)=5.7; P=0.0002) (**Figure 2B** right).

To verify the robustness of this result in the context of different reward contingencies, we analysed learning rates in each task condition separately. In both experiments, our task included three different conditions (**Figure 3A**): a "Symmetric" condition, in which both options were associated with a 50% chance of getting a reward; an "Asymmetric" condition, in which one option was associated with a 75% chance of getting a reward, whereas the other option was associated with only a 25% chance; a "Reversal" condition, in which one option was initially associated with a 83% chance of getting a reward and the other option was associated with a 17% chance of getting a reward, but after 12 trials the contingency reversed. We submitted Experiment 1 factual learning rates to a repeated-measure ANOVA with prediction error valence (positive and negative) and task condition as within-subjects factors (**Figure 3B**). Confirming the aggregate result, the ANOVA showed a significant main effect of valence (F(1,19)=26.4, P=5.8e-5), but no effect of condition (F(2,38)=0.7, P>0.5) and, crucially, no

valence by condition interaction (F(2,38)=0.8, P>0.4). We submitted Experiment 2 factual and counterfactual learning rates to a repeated-measure ANOVA with prediction error valence (positive and negative), prediction error type (factual or counterfactual) and condition (Symmetric, Asymmetric and Reversal) as within-subjects factors (**Figure 3C**). Confirming the aggregate result, the ANOVA showed no effect of valence (F(1,19)=0.0, P>0.9), no effect of type (F(1,19)=0.3, P>0.5), but a significant valence by type interaction (F(1,19)=162.9, P=9.1e-11). We also found an effect of condition (F(2,38)=5.1, P=0.01), reflecting lower average learning rates in the Reversal compared to the Asymmetric condition (T(19)=2.99; P=0.007), which was not modulated by valence (F(2,38)=0.2, P>0.7), or type (F(2,38)=1.2, P>0.3). Crucially, the three-way interaction was not significant (F(2,38)=1.8, P>.1). These results indicate that learning biases were robust across different task contingencies.

Behavioural signatures of learning biases

To investigate the behavioural consequences of the learning biases we median-split all participants according their normalised learning rate differences. We reasoned that the effects of learning biases on behavioural performance could be highlighted comparing participants who differed in the extent they expressed the bias itself. Experiment 1 participants were split according to their normalised factual learning rate bias: $(\alpha_c^+ - \alpha_c^-)/(\alpha_c^+ + \alpha_c^-)$, from which we obtained a high (0.76±0.05) and a low bias (0.11±0.14) group. Experiment 2 participants were split according their normalised confirmation learning rate bias: $[(\alpha_c^+ - \alpha_c^-) - (\alpha_u^+ + \alpha_u^-)]/(\alpha_c^+ + \alpha_c^- + \alpha_u^+ + \alpha_u^-)$, from which we also obtained a high (0.72±0.04) and a low bias (0.36±0.04) group.

From the Symmetric condition we extracted preferred choice rate as a dependent variable, which was the choice rate of the most frequently chosen option (i.e. the option/symbol that was chosen more than >50%). We submitted the preferred choice rate to an ANOVA with experiment (1 or 2) and bias level (high and low) as between-subjects factors. The ANOVA showed a significant main effect of bias level (F(1,36)=8.8, P=0.006). There was no significant main effect of experiment (F(1,36)=0.6, P>0.6) and no significant interaction between experiment and bias level (F(1,36)=0.3, P>0.5). The main effect of bias level was driven by higher preferred choice rate in the high, compared to the low bias group in both Experiment 1 (T(18)=1.8 P=0.08) and Experiment 2 (T(18)=2.3 P=0.03). This result suggests that higher biases were associated with an increased tendency to develop a preferred choice, even in the absence of a “correct” option, which naturally emerges from overweighting positive outcomes[9].

From the remaining conditions we extracted the correct choice rate, which was the choice rate of the most frequently rewarded option. In the Reversal condition, correct choice rate was split across the first (i.e., before the reversal of the contingencies) and second half (i.e., after the reversal of the contingencies) of the trial. We submitted the correct choice rate to a mixed ANOVA with experiment (1 or 2) and Bias Group (high and low) as between-subjects factors, and condition (Asymmetric, Reversal: first half, and Reversal: second half) as a within-subjects factor. We found a main effect of experiment (F(1,36)=4.1, P=0.05), indicating that correct choice rate was higher in Experiment 2 than

Experiment 1, which is consistent with previous studies showing that counterfactual feedback enhances learning[18,21]. We also found a significant effect of bias level ($F(1,36)=10.8$, $P=0.002$), a significant effect of condition ($F(2,72)=99.5$, $P=2.0e-16$), and a significant bias level by condition interaction ($F(2,72)=9.6$, $P=0.0002$). Indeed, in both experiments, the correct choice rate in the second half of the Reversal condition was lower in the high bias compared to the low bias group (Experiment 1: $T(18)=3.9$ $P=0.0003$; Experiment 2: $T(18)=2.5$ $P=0.02$). This result derives from the fact that in the first half of the Reversal condition learning is primarily driven by positive factual prediction errors (and negative counterfactual prediction errors, where applicable), whereas in the second half of the Reversal condition correct performance depends on un-learning previous associations, based on negative factual prediction errors (and positive counterfactual prediction errors, in Experiment 2).

Discussion

Two cohorts of healthy adult participants performed two variants of an instrumental learning task, involving factual (Experiment 1) and counterfactual (Experiments 1 & 2) reinforcement learning. We found that prediction error valence biased factual and counterfactual learning in opposite directions. When learning from obtained outcomes (factual learning), the learning rate for positive prediction errors was higher than the learning rate for negative prediction errors. When learning from forgone outcomes (counterfactual learning), the learning rate for positive prediction errors was lower than that of negative prediction errors. This result proved stable across different reward contingency conditions. Finally, model-free analysis showed that participants with a higher valence-induced learning bias displayed poorer learning performance, specifically when it was necessary to adjust their behaviour in response to a reversal of reward contingencies.

Our results demonstrating a factual learning bias replicate previous findings showing that in simple instrumental learning tasks, participants preferentially learn from positive compared to negative prediction errors[10–12]. However, in contrast to previous studies in which this learning bias had no negative impact on behavioural performance (i.e., correct choice rate and therefore final payoff), here we demonstrated that this learning bias is still present in situations where it has a negative impact on performance. In fact, whereas low and high bias participants performed equally well in conditions with stable reward contingencies, in conditions with unstable reward contingencies we found that high bias participants showed a relatively reduced correct choice rate. In the Reversal condition, learning to successfully reverse the response in the second half of the trials is mainly driven by negative factual (and positive counterfactual) prediction errors, however participants displaying higher biases presented a lower correct choice rate, specifically in the second half of the “Reversal” condition.

In addition to reduced reversal learning, and in accordance with a previous study[9], another behavioural feature that distinguished higher and lower bias participants was the preferred response rate in the Symmetric condition. In the Symmetric condition, both cues had the same reward probabilities (50%), such that there was no intrinsic “correct” response, allowing us to calculate a “preferred” response rate for each participant (defined as the choice rate of the option most frequently selected by a given participant, i.e. the option selected in > 50% of trials). The preferred response rate can therefore be taken as a measure of the tendency to overestimate the value of one cue compared to the other, in the absence of actual outcome-based, factual evidence. In both experiments, higher bias participants showed higher preferred response rates, a behavioural pattern that is consistent with an increased tendency to discount negative factual (and positive counterfactual) prediction errors, which can result in one considering a previously rewarded chosen option as better than it really is and an increased preference for this choice.

Previous studies have so far been unable to distinguish whether this valence-induced factual learning bias was a valuation or a confirmation bias. In other words, do participants preferentially learn from positive prediction errors because they are positively valenced or because the outcome “confirms” the

choice they have just made? To address this question we designed Experiment 2, where, by the inclusion of counterfactual feedback, we were able to separate the influence of valence (positive vs. negative) from the influence of choice (chosen vs. unchosen). Crucially, whereas the two competing hypotheses ("valuation" vs. "confirmation") predict the same result concerning factual leaning rates, they predict opposite effects of valence on counterfactual learning rates. The results from Experiment 2 support the "confirmation" bias hypothesis: participants preferentially took into account the outcomes that "confirmed" their current behavioural policy (positive chosen and negative unchosen outcomes) and discounted the outcomes that contradicted it (negative chosen and positive unchosen outcomes). Our results therefore support the idea that confirmation biases are pervasive in human cognition[22]. It should be noted that from an orthodox Bayesian perspective a "confirmation bias" would involve reinforcing one's own pre-exiting beliefs or preferences. Here, we take a slightly different perspective by extending the notion to reinforcing one's own current choice.

We performed our learning rate analysis separately for each task condition and the results proved robust and not driven by any particular reward contingency condition. While our results contrast with previous studies that have found learning rates adapted as a function of task contingencies, showing increases when task contingencies were unstable[23,24], several differences between these tasks and ours may explain this discrepancy. First, in previous studies the stable and unstable phases were clearly separated, whereas in our design participants were simultaneously tested with the three reward contingency conditions. Second, in our experiments we did not explicitly tell participants to monitor the stability of the reward contingency. Finally, since in our task the Reversal condition represented only fourth quarter of the trials, participants may not have explicitly realized that changing learning rates were adaptive in some cases.

Why do these learning biases have an overall maladaptive value? One possibility is that these learning biases are maladaptive, but they arise from neurobiological constraints, which limit human learning capacity. However, we believe this interpretation is unlikely because we see no clear reason why such limits would differentially affect learning from positive and negative prediction errors. In other words, we would predict that a neurobiological constraint on learning rate would limit all learning rates in a similar way and therefore not produce valence-induced learning asymmetries.

A second possibility is that these learning biases are not maladaptive. For instance it has been shown that in certain reward conditions agents displaying valence-induced learning bias may outperform unbiased agents [25]. Thus, a possible explanation for these learning biases is that they have been positively selected because they can be adaptive in the context of the natural environment in which the learning system evolved[26].

Finally, a third, intermediate possibility is that these learning biases can be maladaptive in the context of learning performance, but due to their adaptive effects in other domains of cognition, overall they have a net adaptive value. For example, these biases may also manifest as "self-serving", choice-

supportive biases, which result in individuals tending to ascribe success to their own abilities and efforts, but relatively tending to neglect failures[27]. These psychological processes may help nourish self-esteem and confidence, both of which have been associated with overall favourable real life outcomes[28].

To conclude, by showing that both factual and counterfactual learning are affected by valence-induced biases, our study highlights the importance of investigating reinforcement learning biases. Most of the decisions we face everyday are experience-based, therefore increasing our understanding of learning biases will likely enable the refinement of existing models of value-based decision-making, furthering our understanding of human cognition[3].

Methods

Participants

The study included two experiments. Each experiment involved N=20 participants (Experiment 1: 7 males, mean age 23.9 ± 0.7; Experiment 2: 4 males, mean age 22.8 ± 0.7). The local ethics committee approved the study. All participants gave written informed consent before inclusion in the study, which was carried out in accordance with the declaration of Helsinki (1964, revised 2013). The inclusion criteria were being older than 18 years and reporting no history of neurological or psychiatric disorders.

Behavioural tasks

Participants performed a probabilistic instrumental learning task based on previous studies[17,18] (Fig. 1A & 1B, Fig. 3A). Briefly, the task involved choosing between two cues that were presented in fixed pairs and therefore represented fixed choice contexts. Cues were associated with stationary outcome probability in three out of four contexts. In the remaining context outcome probability was non-stationary. The possible outcomes were either winning or losing a point. To allow learning, each context was presented 24 trials. Each session comprised the four learning contexts and therefore included 96 trials. The whole experiment involved two sessions, each including the same number of contexts and conditions, but a different set of stimuli. Thus, the total experiment included 192 trials. The four learning contexts (i.e. fixed pairs of cues) were divided in three conditions. In the “Symmetric” condition each cue was associated with a .50 probability of winning one point. In the “Asymmetric” condition one cue was associated with a .75 probability of winning a point and the other cue was associated with a .25 probability of winning a point. The Asymmetric condition was implemented in two choice contexts in each session. Finally, in the “Reversal” condition one cue was associated with a .83 probability of winning a point and the other cue was associated with a .17 probability of winning a point, during the first 12 trials, and these contingencies were reversed thereafter. We chose a bigger probability difference in the Reversal compared to the Asymmetric condition in order to ensure that participants were able to reach a plateau within the first 12 trials. Participants were encouraged to accumulate as many points as possible and were informed that some cues would result in winning more often than others. Participants were given no explicit information regarding reward probabilities, which they had to learn through trial and error.

At each trial, after a fixation cross, the choice context was presented. Participants made their choice by pressing left or right arrow keys with their right hand (the choice time was self-paced). The two experiments differed in the fact that in the Experiment 1 participants were only informed about the outcome of their own choice (chosen outcome), whereas in the Experiment 2 participants were informed about both the obtained and the forgone outcome (i.e. counterfactual feedback). In Experiment 1 positive outcomes were presented on the top and negative outcome of the bottom of the screen. The participant was required to press the key corresponding to the position of the outcome on the screen in order to move to the subsequent trial (top/bottom). In Experiment 2 the obtained outcomes were presented in the same place of the chosen cue and the forgone outcome in the same

place of the unchosen cue. To move to the subsequent trial, participants had to match the position of the outcome (right/left). Importantly for our computational analyses, outcome probabilities (although on average anti-correlated in the Asymmetric" and Reversal conditions) were truly independent across cues, so that in the Symmetric condition, in a given trial, the obtained and forgone outcomes were the same in the 50% of the trials; in the Asymmetric condition this was the case in the 37.5% of the trials; finally, in the Reversal condition this was the case in the 28.2% of the trials.

Behavioural variables

We extracted the correct response rate, that is, the rate of the trials in which the participants chose the most rewarding response, from the Asymmetric and the Reversal conditions. The correct response rate in the Reversal condition was calculated separately for the two phases: before ("first half") and after ("second half") the contingency reversal. In the Symmetric condition, we calculated the so-called "preferred" response rate. The preferred response was defined as the most frequently chosen option, i.e. that chosen by the participant on more than 50% of the trials. This quantity is therefore, by definition, greater than 0.5.

Computational Models

We fitted the data with a standard Q-learning model, including different learning rates following positive and negative prediction errors and containing two different modules (Fig 1.C): a factual learning module to learn from chosen outcomes (R_c) and a counterfactual learning module to learn from unchosen outcomes (R_u) (note that counterfactual learning applies only to Experiment 2). For each pair of cues (choice context), the model estimates the expected values of the two options (Q-values). These *Q*-values essentially represent the expected reward obtained by choosing a particular option in a given context. In both experiments, Q-values were set at 0 before learning, corresponding to the a priori expectation of 50% chance of winning 1 point, plus a 50% chance of losing 1 point. After every trial *t*, the value of the chosen option is updated according to the following rule (factual learning module):

(1)

$$Q_c(t+1) = Q_c(t) + \begin{matrix} \alpha_c^+ . PE_c(t) & \quad if\ PE_c(t) > 0 \\ \alpha_c^- . PE_c(t) & \quad if\ PE_c(t) < 0 \end{matrix}$$

In this first equation, $PE_c(t)$ is the prediction error of the chosen option, calculated as:

(2)

$$PE_c(t) = R_c(t) - Q_c(t),$$

where $R_c(t)$ was the reward obtained as an outcome of choosing c at trial t. In other words, the prediction error $PE_c(t)$ is the difference between the expected outcome $Q_c(t)$ and the actual outcome $R_c(t)$.

In Experiment 2 the unchosen option value was also updated according to following rule (counterfactual learning module):

(3)

$$Q_u(t+1) = Q_u(t) + \begin{matrix} \alpha_u^+ . PE_u(t) & if\ PE_u(t) > 0 \\ \alpha_u^- . PE_u(t) & if\ PE_u(t) < 0 \end{matrix}$$

In this second equation, $PE_u(t)$ is the prediction error of the unchosen option, calculated as:

(4)

$$PE_u(t) = R_u(t) - Q_u(t),$$

where $R_u(t)$ was the reward that could have been obtained as an outcome of having chosen u at trial t. In other words, the prediction error $PE_u(t)$ is the difference between the expected outcome $Q_u(t)$ and the actual outcome $R_u(t)$ of the unchosen option.

The learning rates α_c^+ and α_u^+ are scaling parameters that adjust the amplitude of value changes from one trial to the next when prediction errors of chosen and unchosen option respectively are positive (when the actual reward $R(t)$ is better than the expected reward $Q(t)$) and the learning rates α_c^- and α_u^- do the same when prediction errors are negative. Thus, our model allows for the amplitude of the update to be different following positive and negative prediction errors and for both chosen and unchosen options. It therefore allows for the existence of valence-dependent learning biases. For the sake of model comparison we also considered an unbiased model, where outcome valence does not affect the learning rates $(\alpha_x^+ = \alpha_x^-)$.

Finally, the probability (or likelihood) of selecting the chosen option was estimated with a the soft-max rule as follow:

(5)

$$Pc(t) = e^{(Q_c(t)*\beta)} / (e^{(Q_c(t)*\beta)} + e^{(Q_u(t)*\beta)}).$$

This is a standard stochastic decision rule that calculates the probability of selecting one of a set of options according to their associated values. The temperature, β, is another scaling parameter that adjusts the stochasticity of decision-making.

Model Comparison

In a first analysis, we optimized model parameters by minimizing the negative log-likelihood of the data, given different parameter settings, using Matlab's fmincon function (ranges: 0<β<Infinite, and 0< α_n<1):

(6)

$$LPP = \log\big(P(Data|Model)\big)$$

Negative log-likelihoods (LL) were used to compute at the individual level (random effects) for each model the Bayesian information criterion (BIC), as follows:

(7)

$$BIC = \log(ntrials) * df + 2 * LL$$

BIC where compared between biased and unbiased models to verify that the utilization of the biased model is justified, even accounting for its extra-complexity (**Table 1**).

The parameter optimization procedures

In a second analysis, we optimized model parameters by minimizing the logarithm of the Laplace approximation to the model evidence (or log posterior probability: LPP):

(8)

$$LPP = \log\left(P(Data|Model, Parameters)\right)$$

The LPP increases with the likelihood (a measure of quality of fit) and is penalized by the size of the parameter space (a measure of model complexity). Thus, also the LPP represents a trade-off between accuracy and complexity and can guide model selection. Individual LPPs were fed into the mbb-vb-toolbox[29], a procedure that estimates the expected frequencies and the exceedance probability for each model within a set of models, given the data gathered from all participants. Expected frequency is a quantification of the posterior probability of the model (denoted PP), i.e. the probability of the model generating the data obtained from any randomly selected participant. Exceedance probability (denoted XP) is the probability that a given model fits the data better than all other models in the set, i.e. has the highest PP (**Table 2**).

Given that LPP maximization,(temperature: gamma(1,5); learning rates beta(1.1,1.1)) by including priors over the parameters, avoids degenerate parameter estimates, due to the noisiness of the data, learning rates statistics have been performed value retrieved these analyses. To avoid bias in learning rate comparison, the same priors were used for all learning rates ($\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_c^-$).). In the main analysis, a single set of parameters was used to fit all conditions. In a control analysis, different sets of parameters were used to fit each condition ("Symmetric", "Asymmetric" and "Reversal").

Table 1: Model comparison. BIC: Bayesian Information Criterion; PP: Posterior Probability; XP: Exceedance Probability.

Experiment 1	**BIC**	**PP**	**XP**	**Experiment 2**	**BIC**	**PP**	**XP**
Unbiased 1 (2df)	198.4±11.2	0.24±0.07	0.0	**Unbiased 1 (3df)**	178.2±11.2	0.05±0.03	0.0
Biased 2 (3df)	183.4±12.1	0.74±0.07	1.0	**Biased 2 (5df)**	162.0±13.4	0.95±0.03	1.0

Parameter recovery

To validate our results, and more specifically to verify that valence-induced differences in learning rates reflect true differences in learning, as opposed to an artefact of the parameter optimization procedure, we checked the capacity of recovering the correct parameters in simulated datasets. To do so, we simulated performance on our behavioural task by virtual participants with different learning rates (Fig. S1 & S2). Concerning Experiment 1, we simulated unbiased ($\alpha_c^+ = \alpha_c^-$) and biased ($\alpha_c^+ > \alpha_c^-$) participants. Concerning Experiment 2, we simulated unbiased ($\alpha_c^+ = \alpha_c^-$ and $\alpha_u^+ = \alpha_u^-$), semi-biased ($\alpha_c^+ > \alpha_c^-$ and $\alpha_u^+ = \alpha_u^-$) and biased ($\alpha_c^+ > \alpha_c^-$ and $\alpha_u^+ > \alpha_u^-$) participants. We simulated N=100 virtual participants per phenotype. The results of these analyses are presented in the supplementary materials (Fig. S1 and Fig. S2).

Figures

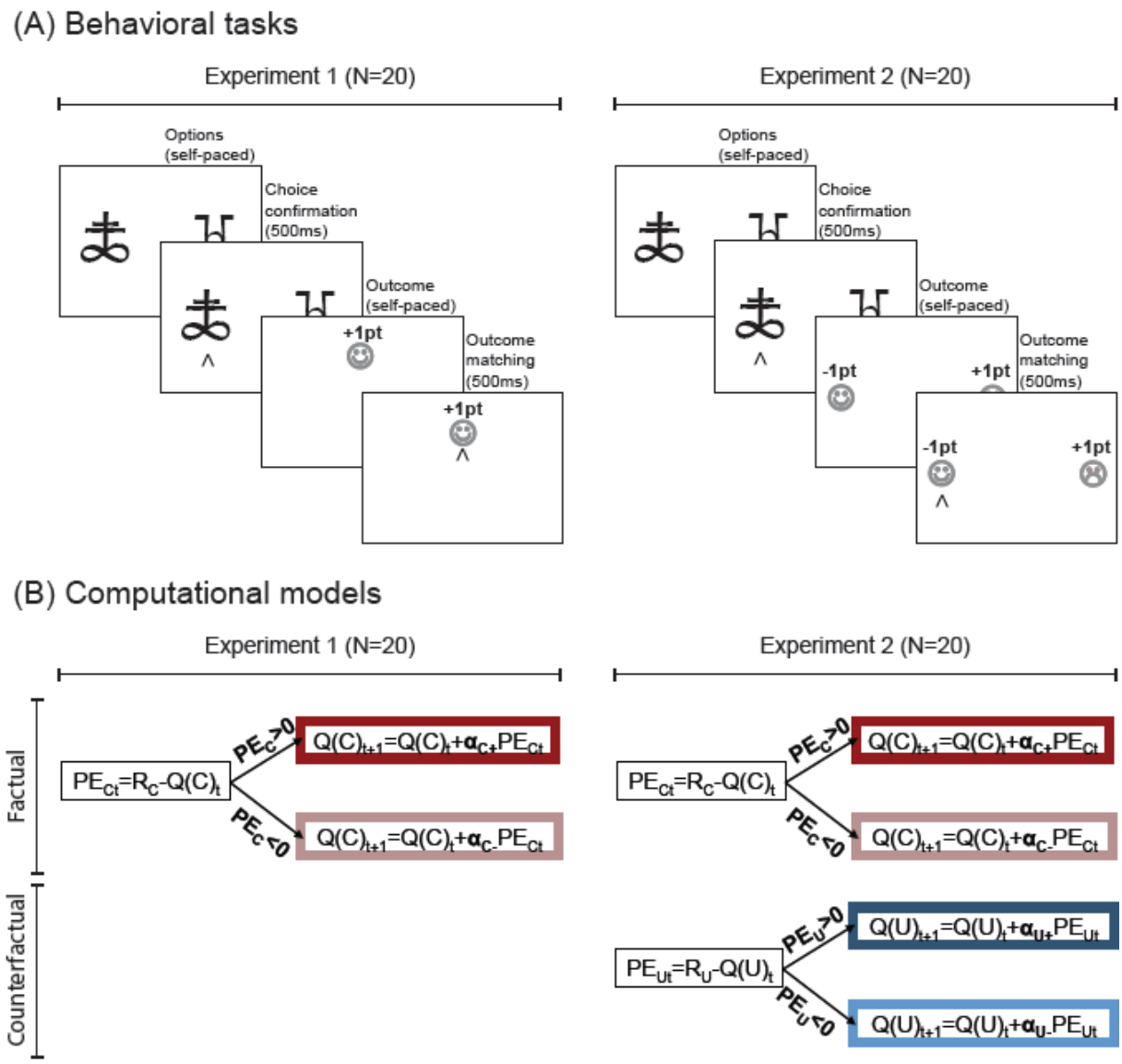


Figure 1: Behavioural task variants and computational model.

(**A**) In Experiment 1 (leftmost panel) participants were shown only the outcome of the chosen option. In Experiment 2 (rightmost panel) participants were shown the outcome of both the chosen and the unchosen options. (**B**) Computational model. The schematic summarizes the value update stage of our computational model. The model contains two computational modules, a factual learning module (in red) to learn from chosen outcomes (R_C) and a counterfactual learning module (in blue) to learn from unchosen outcomes (R_U) (note that the counterfactual learning module does not apply to Experiment 1). Chosen (Q_C) and unchosen (Q_U) option values are updated with delta rules that use different learning rates for positive and negative factual (PE_C) and counterfactual predictions errors (PE_U).

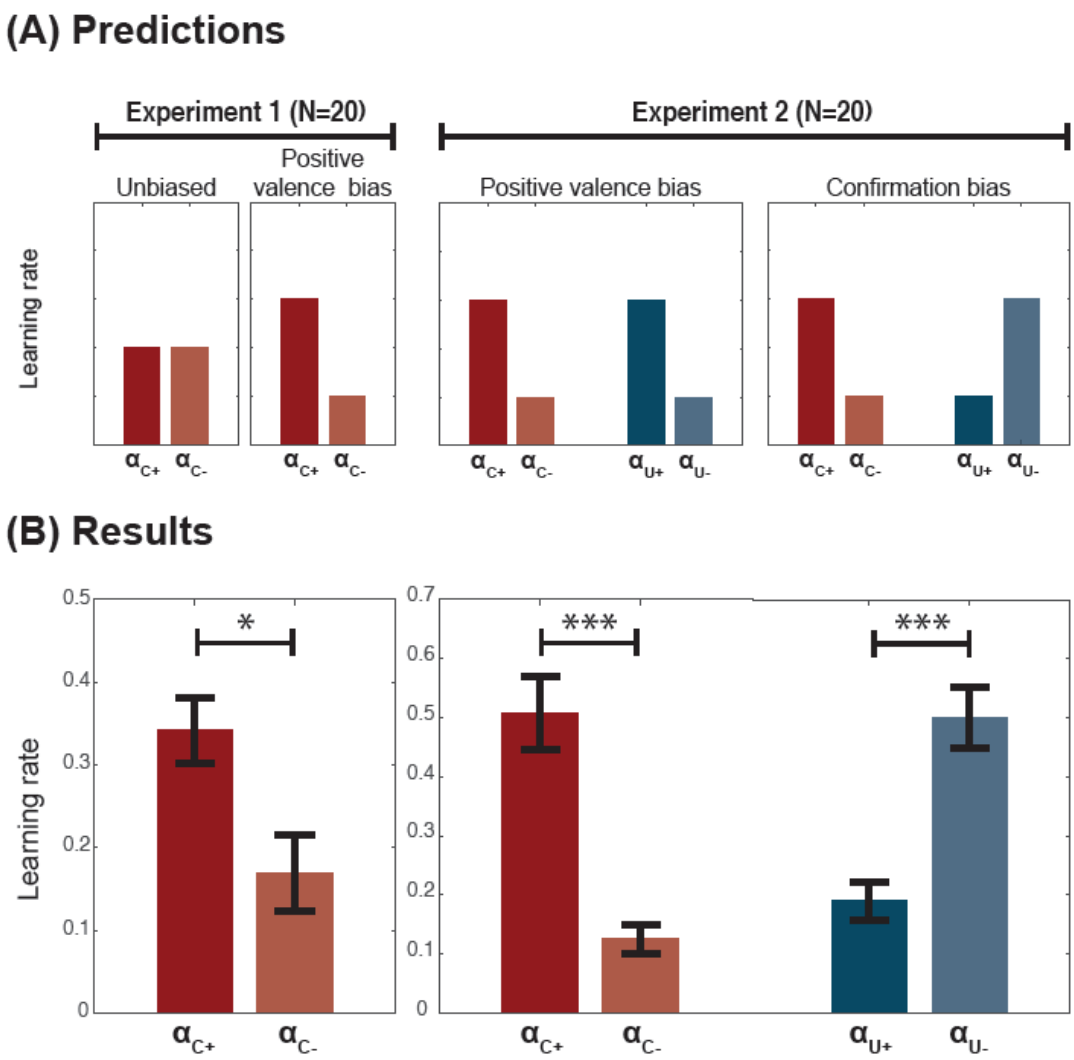


Figure 2: Factual and counterfactual learning biases.

(**A**) **Predicted results.** Based on previous studies we expected that in Experiment 1 factual learning would display a "positive valence" bias (i.e. the learning rate for the chosen positive outcomes would be relatively higher than higher than that of the chosen negative outcomes ($\alpha_c^+ > \alpha_c^-$). In Experiment 2, one possibility was that this "positive valence" bias would extend to counterfactual learning, whereby positive outcomes are over-weighted regardless of whether the outcome was chosen or unchosen ("valuation" bias) ($\alpha_u^+ > \alpha_u^-$). Another possibility was that counterfactual learning would present an opposite bias, whereby the learning rate for the unchosen negative outcomes was higher than the learning rate of the unchosen positive outcomes ($\alpha_u^+ < \alpha_u^-$) ("confirmation" bias). (**B**) **Actual results.** Learning rate analysis of Experiment 1 replicated previous findings, demonstrating that factual learning presents a "positive valence", or "optimistic" bias. Learning rate analysis of Experiment 2 indicated that counterfactual learning was also biased, in a direction that was consistent with a "confirmation" bias. ***P<0.001 and *P<0.05, two-tailed paired t-test.

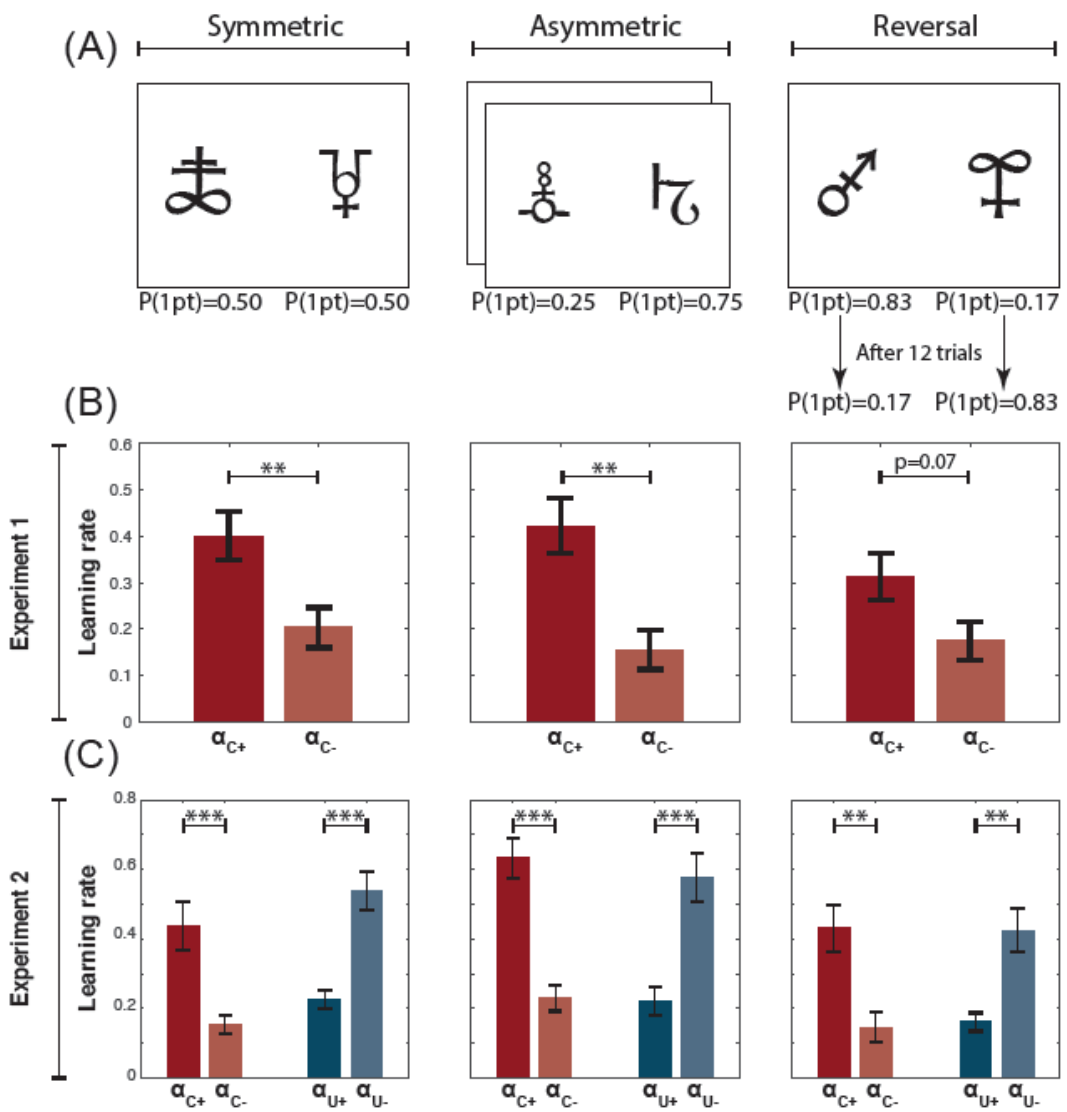


Figure 3: Stability of learning biases across task conditions.

(**A**) **Task conditions.** The Symmetric condition was characterized by a stable reward contingency and no "correct" option, because the two options have equal reward probabilities. The Asymmetric conditions were also characterized by a stable reward contingency and a "correct" option, since one option had a higher reward probability than the other. The Reversal condition was characterized by an instable reward contingency: after 12 trials the reward probability reversed across symbols, so that the former "correct" option became the "incorrect" one, and vice versa. Note that the number of trials refers to one session and participants performed two sessions, each involving new pairs of stimuli (192 trials in total). (**B**) and (**C**) Computational results as a function of the task conditions in Experiment 1 and Experiment 2, respectively. Each column presents the result of the corresponding condition presented in (**A**). ***P<0.001 and **P<0.01, two-tailed paired t-test.

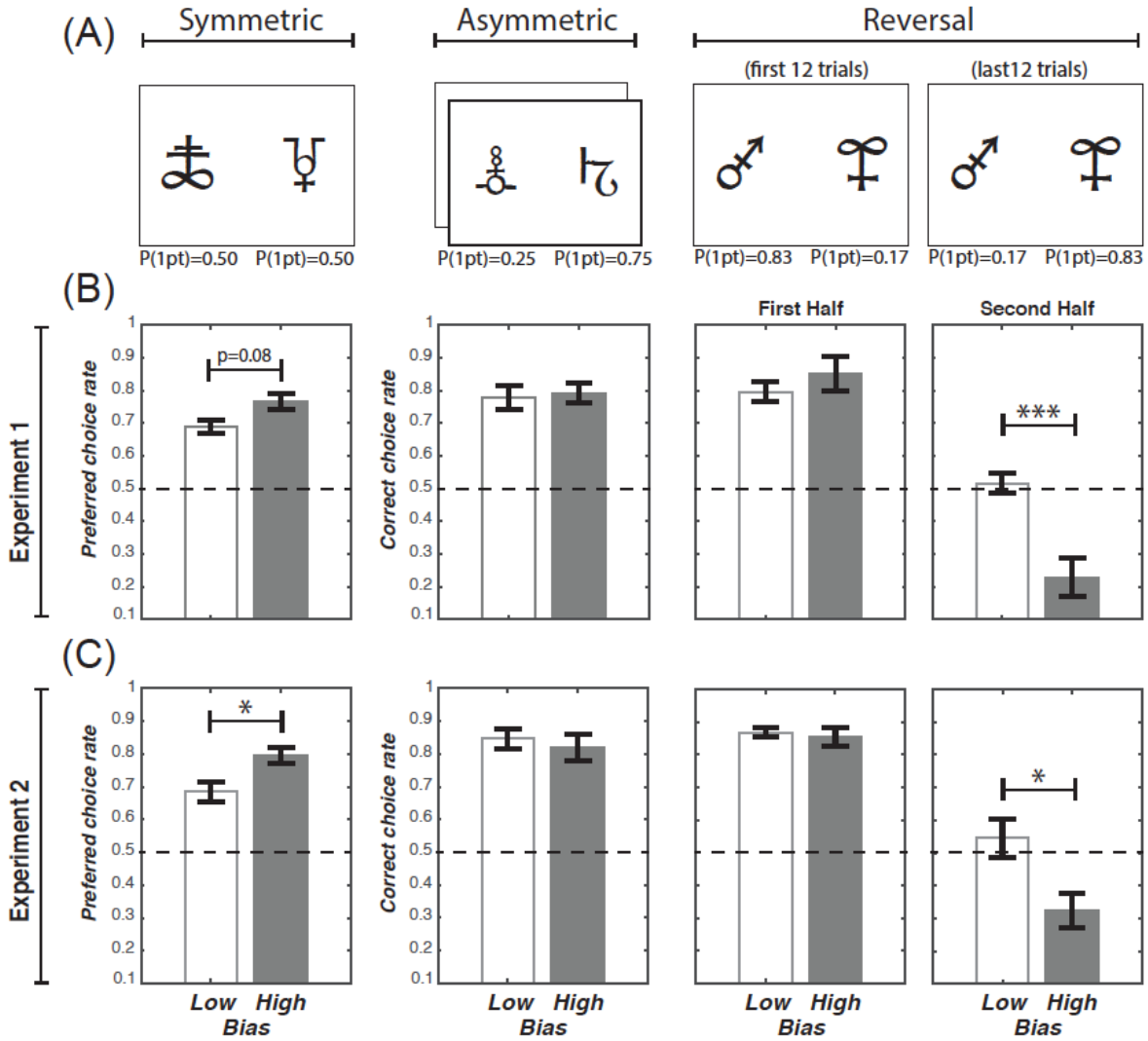
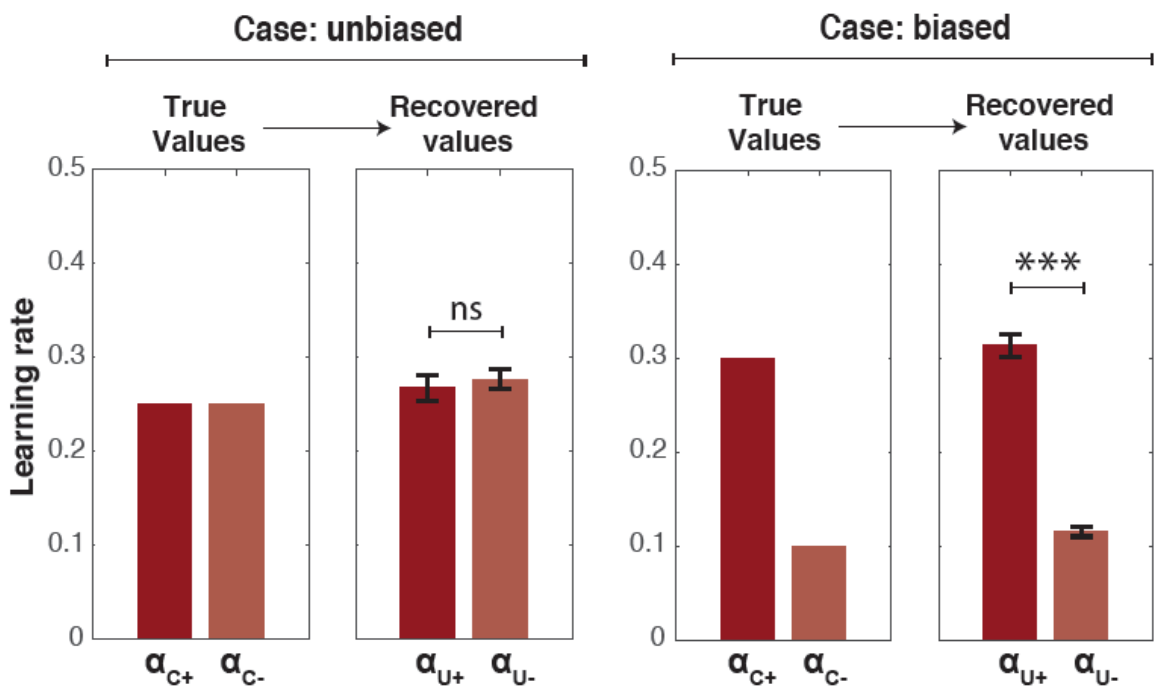


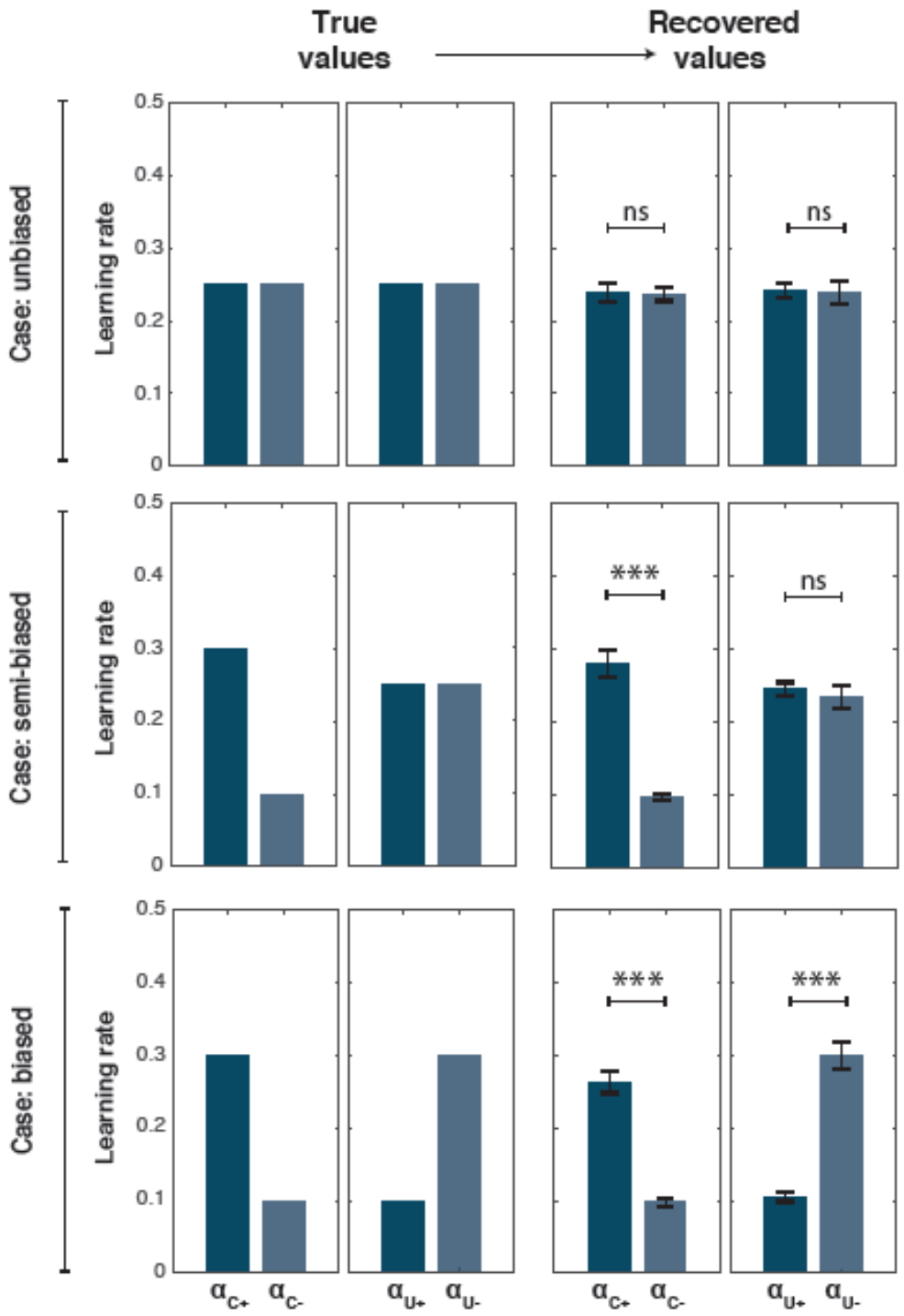
Figure 4: Behavioral signatures distinguishing "low" and "high bias" participants**.**

(**A**) Task conditions as in **Figure 3A**. (**B**) and (**C**) Behavioural results as a function of the task conditions in Experiment 1 and Experiment 2, respectively. Each column presents the result of the corresponding condition presented in (**A**). In the Symmetric conditions, where there was no correct option, we calculated the "Preferred choice rate", which was the choice rate of the most frequently chosen option (by definition, this was always greater than 0.5). In the Asymmetric and the Reversal conditions we calculated the correct choice rate. In the Reversal condition the correct choice rate was split between the two learning phases. ***P<0.001 and *P<0.05, two-tailed paired t-test.



Supplementary Figure 1: Parameter recovery in the Experiment 1 setting.

"True values": learning rates used to simulate the data. "Recovered values": learning rates obtained from the simulations once the same parameter optimization was applied as for the experimental data. "Case: unbiased": no learning rate bias. "Case: biased": optimistic learning rate bias.



Supplementary Figure 2: parameter recovery in the Experiment 2 setting.

"True values": learning rates used to simulate the data. "Recovered values": learning rates obtained from the simulations once applied the same parameter optimization as for the experimental data. "Case: unbiased": no learning rate bias. "Case: semi-biased": learning rate bias only concerning factual learning. "Case biased": confirmation bias involving both factual and counterfactual learning.

Acknowledgments

Anahit Mkrtchian and Anders Jespersen performed the experiments. We thank Bahador Bahrami and Valerian Chambon for helpful discussions.

Financial disclosure

SP and the study were supported by a Marie Sklodowska-Curie Individual European Fellowship (PIEF-GA-2012 Grant 328822). SP is currently supported by an ATIP-Avenir grant. GL was supported by a PHD fellowship of the Ministère de l'enseignement supérieur et de la recherche. EJK is supported by a Medical Research Council studentship. SJB is funded by a Royal Society University Research Fellowship and the Jacobs foundation. The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Bibliography

1. Rangel A, Camerer C, Montague PR. A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci. 2008;9: 545–556. doi:10.1038/nrn2357
2. DellaVigna S. Psychology and Economics: Evidence from the Field. J Econ Lit. 2009;47: 315–372. doi:10.1257/jel.47.2.315
3. Hertwig R, Erev I. The description-experience gap in risky choice. Trends Cogn Sci. 2009;13: 517–23. doi:10.1016/j.tics.2009.09.004
4. Maia T V, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. Nature Publishing Group; 2011;14: 154–62. doi:10.1038/nn.2723
5. Haushofer J, Fehr E. On the psychology of poverty. Science (80-). 2014;344: 862–867. doi:10.1126/science.1232491
6. Doya K. Metalearning and neuromodulation. Neural Netw. 2002;15: 495–506. Available: http://www.ncbi.nlm.nih.gov/pubmed/12371507
7. Barto AG, Sutton RS. Reinforcement Learning: An Introduction. Cambridge. MIT Press; 1998. doi:10.1109/TNN.1998.712192
8. Friston KJ, Daunizeau J, Kiebel SJ. Reinforcement learning or active inference? PLoS One. 2009;4: e6421. doi:10.1371/journal.pone.0006421
9. Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S. Asymmetric reinforcement learning: computational and neural bases of positive life orientation [Internet]. bioRxiv. Cold Spring Harbor Labs Journals; 2016 Feb. doi:10.1101/038778
10. den Ouden HEM, Daw ND, Fernandez G, Elshout J a, Rijpkema M, Hoogman M, et al. Dissociable effects of dopamine and serotonin on reversal learning. Neuron. 2013;80: 1090–100. doi:10.1016/j.neuron.2013.08.030
11. Frank MJ, Moustafa A a, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 2007;104: 16311–16316. doi:10.1073/pnas.0706111104
12. van den Bos W, Cohen MX, Kahnt T, Crone E a. Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. Cereb Cortex. Oxford University Press; 2012;22: 1247–55. doi:10.1093/cercor/bhr198
13. Sharot T, Garrett N. Forming Beliefs: Why Valence Matters. Trends Cogn Sci. Elsevier Ltd; 2016;20: 25–33. doi:10.1016/j.tics.2015.11.002
14. O'Doherty JP, Lee SW, McNamee D. The structure of reinforcement-learning mechanisms in the human

brain. Curr Opin Behav Sci. Elsevier Ltd; 2015;1: 94–100. doi:10.1016/j.cobeha.2014.10.004

15. Boorman ED, Behrens TE, Rushworth MF. Counterfactual Choice and Learning in a Neural Network Centered on Human Lateral Frontopolar Cortex. PLoS Biol. 2011;9. doi:10.1371/journal.pbio.1001093
16. Fischer AG, Ullsperger M. Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. Neuron. Elsevier Inc.; 2013;79: 1243–55. doi:10.1016/j.neuron.2013.07.006
17. Palminteri S, Boraud T, Lafargue G, Dubois B, Pessiglione M. Brain hemispheres selectively track the expected value of contralateral options. J Neurosci. 2009;29: 13465–13472. doi:10.1523/JNEUROSCI.1500-09.2009
18. Palminteri S, Khamassi M, Joffily M, Coricelli G. Contextual modulation of value signals in reward and punishment learning. Nat Commun. Nature Publishing Group; 2015;6: 8096. doi:10.1038/ncomms9096
19. Rescorla R a., Wagner AR. RescorlaWagnerChapter1972.pdf. Classical conditioning II: current research and theory. 1972. pp. 64–99.
20. Watkins CJCH, Dayan P. Q-learning. Mach Learn. 1992;8: 279–292. doi:10.1007/BF00992698
21. Palminteri S, Kilford EJ, Coricelli G, Blakemore S-J. The computational development of reinforcement learning during adolescence. PLoS Comput Biol. 2016;
22. Nickerson R. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol. 1998;2: 175–220.
23. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci. 2007;10: 1214–21. doi:10.1038/nn1954
24. Browning M, Behrens TE, Jocham G, Reilly JXO, Bishop SJ. Anxious individuals have difficulty learning the causal statistics of aversive environments Michael Browning. Nat Publ Gr. Nature Publishing Group; 2015;18: 1–50. doi:10.1038/nn.3961
25. Cazé RD, van der Meer MAA. Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern. 2013;107: 711–719. doi:10.1007/s00422-013-0571-5
26. Fawcett TW, Fallenstein B, Higginson AD, Houston AI, Mallpress DEW, Trimmer PC, et al. The evolution of decision rules in complex environments. Trends Cogn Sci. Elsevier Ltd; 2014;18: 153–161. doi:10.1016/j.tics.2013.12.012
27. Blaine B, Crocker J. Self-Esteem and Self-Serving Biases in Reactions to Positive and Negative Events: An Integrative Review. In: Baumeister RF, editor. Self-Esteem. The Springer Series in Social Clinical Psychology; 1993. p. pp 55-85.
28. Weinstein ND. Unrealistic Optimism About Future Life events. J Pers Soc Psychol. 1980;39: 806–820. doi:10.1037/a0020997
29. Daunizeau J, Adam V, Rigoux L. VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLoS Comput Biol. 2014;10: e1003441. doi:10.1371/journal.pcbi.1003441