Published in final edited form as:
Neural Comput. 2022 January 14; 34(2): 307–337. doi:10.1162/neco_a_01455.

A normative account of confirmation bias during reinforcement learning

Germain Lefebvre[1], **Christopher Summerfield**[#2], **Rafal Bogacz**[#1,*]

[1]MRC Brain Network Dynamics Unit, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

[2]Department of Experimental Psychology, University of Oxford, Oxford, UK

[#] These authors contributed equally to this work.

Abstract

Reinforcement learning involves updating estimates of the value of states and actions on the basis of experience. Previous work has shown that in humans, reinforcement learning exhibits a confirmatory bias: when updating the value of a chosen option, estimates are revised more radically following positive than negative reward prediction errors, but the converse is observed when updating the unchosen option value estimate. Here, we simulate performance on a multi-arm bandit task to examine the consequences of a confirmatory bias for reward harvesting. We report a paradoxical finding: that confirmatory biases allow the agent to maximise reward relative to an unbiased updating rule. This principle holds over a wide range of experimental settings and is most influential when decisions are corrupted by noise. We show that this occurs because on average, confirmatory biases lead to overestimating the value of more valuable bandits, and underestimating the value of less valuable bandits, rendering decisions overall more robust in the face of noise. Our results show how apparently suboptimal learning rules can in fact be reward-maximising if decisions are made with finite computational precision.

Introduction

Confirmation bias refers to seeking or interpreting evidence in ways that are influenced by existing beliefs, and it is a ubiquitous feature of human perceptual, cognitive and social processes, and a longstanding topic of study in psychology (Nickerson, 1998). Confirmatory biases can be pernicious in applied settings, for example when clinicians overlook the correct diagnosis after forming a strong initial impression of a patient (Groopman, 2007). In laboratory, confirmation bias has been studied with a variety of paradigms (Nickerson, 1998; Talluri, Urai, Tsetsos, Usher, & Donner, 2018). One paradigm in which the confirmation bias can be observed and measured involves reinforcement learning tasks, where participants have to learn from positive or negative feedback which options are worth taking (Chambon et al., 2020; Palminteri, Lefebvre, Kilford, & Blakemore, 2017), and this paper focusses on confirmation bias during reinforcement learning.

[*]Corresponding Author: rafal.bogacz@ndcn.ox.ac.uk.

Europe PMC Funders Author Manuscripts

In the laboratory, reinforcement learning is often studied via a “multi-armed bandit” task in which participants choose between two or more states that pay out a reward with unknown probability (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). Reinforcement learning on this task can be modelled with a simple principle known as a delta rule (Rescorla & Wagner, 1972), in which the estimated value V_t^i of the chosen bandit i on trial t is updated according to:

$$V_{t+1}^i = V_t^i + \alpha \cdot \delta_t^i \quad (1)$$

In this equation α is a learning rate in unity range, and δ_t^i is the reward prediction error defined as

$$\delta_t^i = R_t^i - V_t^i \quad (2)$$

In the above equation, R_t^i is the payout for option i on trial t. If α is sufficiently small, V_t^i tend to converge over time to the vicinity of expected value of bandit i (in stationary environments).

This task and modelling framework have also been used to study the biases that humans exhibit during learning. One line of research has suggested that humans may learn differently from positive and negative outcomes. For example, variants of the model above which include distinct learning rates for positive and negative updates to V_t^i have been observed to fit human data from a 2-armed bandit task better, even after penalising for additional complexity (Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012). Similar differences in learning rates after positive and negative feedback have also been observed in monkeys (Farashahi, Donahue, Hayden, Lee, & Soltani, 2019) and rodents (Cieślak, Ahn, Bogacz, & Parkitna, 2018), suggesting that they reflect an important optimization of a learning process that occurred earlier in evolution and has been preserved across species. When payout is observed only for the option that was chosen, updates seem to be larger when the participant is positively rather than negatively surprised, which might be interpreted as a form of optimistic learning (Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017). However, a different pattern of data was observed in follow-up studies in which counterfactual feedback was also offered – i.e., the participants were able to view the payout associated with both chosen and unchosen options. Following a feedback on the unchosen option, larger updates were observed for negative prediction errors (Chambon et al., 2020; Palminteri et al., 2017; Schuller et al., 2020). This is consistent with a confirmatory bias rather than a strictly optimistic bias, whereby belief revision helps to strengthen rather than weaken existing preconceptions about which option may be better.

One obvious question is why confirmatory biases persist as a feature of our cognitive landscape – if they promote suboptimal choices, why have they not been selected away by evolution? One variant of the confirmation bias, a tendency to overtly sample information from the environment that is consistent with existing beliefs, has been argued to promote

optimal data selection: where the agent chooses its own information acquisition policy, exhaustively ruling out explanations (however obscure) for an observation would be highly inefficient (Oaksford & Chater, 2003). However, this account is unsuited to explaining the differential updates to chosen and unchosen options in a bandit task with counterfactual feedback, because in this case feedback for both options is freely displayed to the participant, and there is no overt data selection problem.

It has been demonstrated previously that biased estimates of value can paradoxically be beneficial in two-armed tasks in the sense that under standard assumptions, they maximise the average total reward for the agent (Caze & van der Meer, 2013). This happens because with such biased value estimates, the difference $V_t^1 - V_t^2$ may be magnified, so with a noisy choice rule (typically used in reinforcement learning models) the option with the higher reward probability is more likely to be selected. Caze and van der Meer (2013) considered a standard reinforcement learning task in which feedback is provided only for the chosen option. In that task the reward probabilities of the two options in the task determine whether it is beneficial to have higher learning rate after positive or negative prediction error (Caze & van der Meer, 2013). In other words, when only outcome of chosen option is observed, optimistic bias is beneficial for some reward probabilities, while pessimistic bias for other.

In this paper we show that if the participants are able to view the payouts associated with both chosen and unchosen options, reward is typically maximized if the learning rates follow the pattern of the confirmation bias, i.e. are higher when the chosen option is rewarded and unchosen option is unrewarded. We find that this benefit holds over a wide range of settings, including both stationary and nonstationary bandits, with different reward probabilities, across different epoch lengths, and under different levels of choice variability. We also demonstrate that such confirmation bias tends to magnify the difference $V_t^1 - V_t^2$ and hence makes the choice more robust to the decision noise. These findings may explain why humans tend to revise beliefs to a smaller extent when outcomes do not match with their expectations.

Below we formalize the confirmation bias in a reinforcement learning model, compare its performance in simulations with models without confirmation bias, and formally characterise the biases introduced in value estimates. We also point that the confirmation bias not only typically increases the average reward, but may shorten reaction times and thus increase the rate of obtaining rewards to even higher extent.

Reinforcement learning models

Confirmation model

We analyse properties of a *confirmation* model (Palminteri et al., 2017), which describes learning in a two-armed bandit task where feedback is provided for both options on each trial. The model updates the corresponding value estimates V_t^i according to a delta rule with two learning rates: α^C for confirmatory updates (i.e. following positive prediction errors for the chosen option, and negative for the unchosen option) and α^D for disconfirmatory updates (i.e. following negative prediction errors for the chosen option, and positive for the unchosen

option) (Palminteri et al., 2017). Thus on each trial t, if the agent chooses the option 1, the model updates the values V_t^1 and V_t^2 of the chosen and unchosen options respectively, such that:

$$V_{t+1}^1 = V_t^1 + \begin{cases} \alpha^C \cdot \delta_t^1, & \text{if } \delta_t^1 > 0 \\ \alpha^D \cdot \delta_t^1, & \text{if } \delta_t^1 < 0 \end{cases} \tag{3}$$

and

$$V_{t+1}^2 = V_t^2 + \begin{cases} \alpha^D \cdot \delta_t^2, & \text{if } \delta_t^2 > 0 \\ \alpha^C \cdot \delta_t^2, & \text{if } \delta_t^2 < 0 \end{cases} \tag{4}$$

with δ_t^i being the prediction error for bandit i on trial t defined in Equation 2. We define an agent with a confirmatory bias as one for whom $\alpha^C > \alpha^D$, whereas an agent with a disconfirmatory bias has $\alpha^C < \alpha^D$, and an agent with no bias (or a neutral setting) has $\alpha^C = \alpha^D$. Note that for $\alpha^C = \alpha^D = \alpha$, the model amounts to a standard delta-rule model with a unique learning rate α defined in Equation 1, and we refer to it as *unbiased*.

Decaying learning rate model

We compared the performance of the *confirmation* model in a stable environment to an optimal value estimator which for each option computes the average of rewards seen so far. Such values can be learned by a model using the update given in Equation 1 with the learning rate α decreasing over trials according to $\alpha = \frac{1}{t}$, where t is the trial number (note that with the counterfactual feedback, t is also equal to the number of times the reward for this option has been observed).

Decision Policies

In this paper we consider three policies for making a choice on the basis of learned values: *hardmax*, *softmax* and *ε-greedy* policies. The hardmax is a noiseless policy selecting deterministically the arm associated to the highest value. The softmax is a probabilistic action selection process associating to each arm a the probability P_t^a of being selected based on their respective values such that:

$$P_t^a = \frac{\exp\left(V_t^a/\beta\right)}{\exp\left(V_t^1/\beta\right) + \exp\left(V_t^2/\beta\right)} \tag{5}$$

In the above equation β is the temperature of the *softmax* function, and the higher the temperature, the more random the decision is. To mathematically analyse properties of the confirmation model, we also consider a simpler stochastic choice rule ε-greedy, which on majority of trials selects option with highest estimated value, while with certain fixed probability selects an action randomly.

Effects of confirmation bias on average reward

Methods of simulation

Our goal was to test how outcomes vary with a confirmatory, disconfirmatory or neutral bias across a wide range of different settings that have been the subject of previous empirical investigation in humans and other animals. We considered tasks involving choice between two options. Each bandit i may yield reward $R = 1$ with probability p^i, and no reward ($R = 0$) with probability $1 - p^i$. Importantly we assumed that the agent observes on each trial the payouts for both options: the chosen one and the not chosen option (Fig. 1a). We consider an agent who chooses among bandits for 2^n trials, where n varied from 2 to 10 in simulations (Fig. 1b), and the presented rewards were averaged over these values of n (unless otherwise stated).

We considered four different ways in which the reward probabilities p^i are set, illustrated schematically in Fig. 1c. First we considered *stable* environments in which reward probabilities were constant. We also considered *1 reversal* and *3 reversals* conditions where the payout probabilities were reversed to $1 - p^i$ once in the middle of the task (second display in Fig. 1c), or three times at equal intervals (third display in Fig. 1c). In *stable*, *1 reversal* and *3 reversals* conditions, the initial probabilities p^i at the start of the task were sampled at intervals of 0.1 in the range [0.05, 0.95] such that $p^1 \neq p^2$, and we tested all possible combinations of these probabilities (that is *45* probability pairs). Unless otherwise noted, results are averaged across these initial probabilities.

Additionally, we considered *random walk* condition where the probabilities were initialized to a random number from uniform distribution on interval from 0 to 1, and then drifted over trials as follows:

$$p^i_{t+1} = p^i_t + \kappa \cdot \left(0.5 - p^i_t\right) + \mathcal{N}\left(0, \sigma^2\right) \tag{6}$$

with κ being a parameter decaying the reward probability towards 0.5 (here set to $\kappa = 0.001$) and σ being the standard deviation of the normal distribution from which the fluctuations in probabilities were sampled (here set to $\sigma = 0.02$). Sample probabilities generated from this process are shown in the fourth display in Fig. 1c.

We conduct all simulations numerically, sampling the initial payout probabilities and experiment length(s) exhaustively, varying α^C and α^D exhaustively, and noting the average reward obtained by the agent in each setting. The model is simulated with all possible combinations of learning rates α^C and α^D defined in the range [0.05,0.95] with increments of 0.05, that is 19^2 learning rate combinations. For each combinations of parameters, the simulations were performed 1000 times for all but the random walk condition where simulations are performed 100000 times to account for the increased variability. Results are averaged for plotting and analysis. In all cases, inferential statistics were conducted using nonparametric tests with an alpha of $p < 0.001$ and Bonferroni correction for multiple comparisons. At the start of each simulation, the value estimates were initialized to $V^i_0 = 0.5$.

Results of simulations

Fig. 2 plots total reward obtained in the stationary bandit problem as a function of α^C (y-axis) and α^D (x-axis), for the sequence length of 1024 and averaged across payout probabilities, for both the hardmax (left) and softmax (right) rules. The key result is that rewards are on average greater when $\alpha^C > \alpha^D$ (warmer colours above the diagonal) relative to when they are equal or when $\alpha^C < \alpha^D$. We tested this finding statistically by repeating our simulations multiple times with resampled stimulus sequences (and choices in the softmax condition) and comparing the accrued reward to a baseline in which $\alpha^C = \alpha^D = 0.05$, i.e. the most promising unbiased setting for α. The area enclosed by black line in Fig. 2a-b indicate combinations of learning rates that yield rewards higher than the unbiased setting. Fig. 2b confirms that in particular for the more plausible case where decisions are noisy (i.e. softmax temperature $\beta > 0$), there is a reliable advantage for a confirmatory update policy in the bandit task.

We compared the performance of the confirmation model to the decaying learning rate model described above, which maximizes reward under the assumption that payout probabilities are stationary and decisions are noiseless (i.e. under a hardmax choice rule). We confirmed this by plotting the average reward under various temperature values for three models: one in which a single learning rate was set to a fixed low value $\alpha = 0.05$ (*small learning rate* model) one in which it was optimally annealed (*decaying learning rate* model), and one in which there was a confirmatory bias (*confirmation* model; Fig. 2c). As can be seen, only under $\beta = 0$ the confirmation bias does not increase rewards; as soon as decision noise increases, the relative merit of the confirmation model grows sharply. Importantly, whereas the performance advantage for the decaying learning rate model in the absence of noise (under $\beta = 0$) was very small (on the order of 0.2%), the converse advantage for the confirmatory bias given noisy decisions was numerically larger (1.6%, 4.6% and 5.5% under $\beta = 0.1, 0.2, 0.3$ respectively).

Next, we verified that these results held over different trial lengths and for differing volatility conditions. The results (averaged over different numbers of trials) are shown in Fig. 3. One can see equivalent results presented for a paradigm involving stable contingencies (Fig. 3a and 3e), a reversal of probability between the two bandits midway through the sequence (Fig. 3b and 3f), for three such reversals (Fig. 3c and 3g), and for a random walk in which probabilities drift upwards or downwards on each trial (Fig. 3d and 3h). When decisions are noisy, in all four cases, confirmatory agents reap more rewards than disconfirmatory agents, and also than agents for whom there is a single α selected to maximise reward (Fig. 3e-h). When the decisions are based on the hardmax choice rule, there was no biased combination of learning rates giving significantly higher rewards than unbiased model (Fig. 3a-d). Nevertheless, there still existed combinations of parameters with $\alpha^C > \alpha^D$ yielding reward similar to that from the unbiased model.

Subsequently, we tested how the sequence length affected the relative advantage conferred by a confirmatory bias. In Fig. 4a, we show that the advantage for the confirmatory over the unbiased model holds true for all but the very shortest sequences and continues to grow up to sequences of 1024 trials. Finally, the confirmatory model is most advantageous at intermediate levels of decisions noise (as quantified here by the softmax temperature). As

Europe PMC Funders Author Manuscripts

we have seen, the relative numerical and statistical advantage is lower if we assume no decision noise, but as decision noise grows to the extent that performance tends towards random, all differences between different update policies disappear (Fig. 4b).

Many decisions faced by humans and animals in natural environments involve choices between multiple options, hence we investigated if the confirmation bias also brings an advantage in such situations. The confirmation model can be naturally extended to multiple options by applying the update of Equation 4 to all unchosen options. Fig. s1 show that confirmation bias also increases the average outcome for extended learning environments with more than two options.

Lastly, we performed simulations of the experiment by Palminteri et al. (2017) in order to see where human participants' learning rates combinations stand in terms of performance. In this study, participants made choices between two options and received feedback on outcomes of both options. The task involved choices in multiple conditions in which the participants could receive outcomes -1 or 1. In some conditions the reward probabilities were constant, while on others 1 reversal occurred. We simulated the confirmation model in the same sets of conditions that were experienced by participants, with the same number of trials. We used the values of softmax temperature estimated from individual participants by fitting the confirmation model to their behaviour (data is available at https://doi.org/10.6084/m9.figshare.4265408.v1). These estimated parameter values of the confirmation model were reported by Palminteri et al. (2017).

Fig. 5 shows for each participant, the simulated performance of all learning rates combinations considering their level of decision noise as observed during the experiment, as well as their fitted learning rates. As expected, most participants' learning rates combinations fall in the vicinity of the best performing learning rates combinations, above the diagonal. This tends to confirm the hypothesis that humans use “biased” learning rates because this increased reward in the presence of noise in the decision process.

Confirmation bias magnifies difference between estimated values

The above simulations show that a confirmatory update strategy – one which privileges the chosen over the unchosen option – is reward-maximising across a wide range of experimental conditions, in particular when decisions are noisy. Why would this be the case? It is well known, for example, that adopting a single small value for α will allow value estimates to converge to their ground truth counterparts. Why would an agent want to learn biased value estimates? To answer this question, we demonstrate below that the confirmation bias often magnifies the differences between estimated values and hence make choices more robust to decision noise. We first show it on an intuitive example and then more formally.

Example of the effects of confirmation bias

We selected three parametrisations of the update rules and examined their consequences in more detail. The selected pairs of values for α^C α^D are illustrated in Fig. 6a (symbols △, × and ○). The first corresponded to an unbiased update rule: $\alpha^C = \alpha^D = 0.25$; the second to a moderately biased rule ($\alpha^C = 0.35$, $\alpha^D = 0.15$); and the third to a severely biased rule (α^C

= 0.45, $\alpha^D = 0.05$). Let us refer to the bandit with a higher reward probability as richer and to the other bandit as poorer. We chose a setting in which reward probability for the richer bandit is $p^+ = 0.65$, while for the poorer bandit it is $p^- = 0.35$.

For each update rule, we plotted the evolution of the value estimate for the richer bandit V^+ over trials (Fig. 6b) as well as aggregate choice accuracy (Fig. 6c). Beginning with the choice accuracy data, one can see that intermediate levels of bias are reward-maximising, in the sense that they increase the probability that the agent chooses the bandit with the higher payout probability, relative to an unbiased or a severely biased update rule (Fig. 6c). This is of course simply a restatement of the finding that biased policies maximise reward (see shading in Fig. 6a). However, perhaps more informative are the value estimates for V^+ under each update rule (Fig. 6b). As expected, the unbiased learning rule allows the agent to accurately learn the appropriate value estimate, such that after a few tens of trials, $V^+ \approx p^+ = 0.65$ (grey line). By contrast, the confirmatory model *overestimates* the value of the richer option (converging close to $V^+ \sim 0.8$ despite $p^+ = 0.65$, and (not shown) the model *underestimates* the value of the poorer option $p^- = 0.35$). Thus, the confirmation model outperforms the unbiased model despite misestimating the value of both the better and the worse option. How is this possible?

To understand this phenomenon, it is useful to consider the policy by which simulated choices are made. In the two-armed bandit case, the softmax choice rule of Equation 5 can be re-arranged to the following logistic function:

$$P_t^1 = \frac{1}{1 + \exp\left(\left(V_t^2 - V_t^1\right)/\beta\right)} \tag{7}$$

Here, the choice probability depends both on the inverse slope of the choice function β and the difference in value estimates for bandits 1 and 2. The effect of the confirmation bias is to inflate the quantity $V_t^1 - V_t^2$ away from zero in either the positive or the negative direction, thereby ensuring choice probabilities that are closer to 0 or 1 even in the presence of decision noise (i.e. larger β) This comes at a potential cost of overestimating the value of the poorer option rather than the richer, which would obviously hurt performance. The relative merits of an unbiased vs. biased update rule are thus shaped by the relative influence of these factors. When the rule is unbiased, the model does not benefit from the robustness conferred by inflated value estimates. When the model is severely biased, the probability of confirming the incorrect belief is excessive – leading to a high probability that the poorer option will be overvalued rather than the richer (see the bimodal distribution of value estimates in Fig. 6b, inset). Our simulations show that when this happens, the average reward is low, resulting in bimodal distribution of rewards across simulations (inset in Fig. 6a). However, there exists a "goldilocks zone" for confirmatory bias in which the benefit of the former factor outweighs the cost of the latter. This is why a confirmation bias can help maximise reward.

Analysis of biases in estimated values

This section shows formally that the confirmation bias tends to increase the distance between the estimated values, but beyond certain critical level of confirmation bias the

model may get stuck in a false belief that the poorer option is superior. We followed the approach from a previous study analysing biases in values due to unequal learning rates (Caze & van der Meer, 2013) and analysed the values learned in a stable environment. Due to stochastic nature of rewards in the task, V_t^i constantly fluctuate, but with time they approach a vicinity of values known as stochastic fixed points, in which they will not change on average, i.e. $\mathrm{E}\left(\Delta V_t^i\right) = 0$ (**E** denotes expected value). The fluctuation of estimated values around stochastic fixed points is illustrated in Fig. 7a. Different displays correspond to different levels of confirmation bias quantified by $b = \frac{\alpha^C}{\alpha^D}$. For relatively low levels of bias there exists only a single fixed point V_{true}^i for each estimated value, corresponding to a true belief that the richer option is superior. Comparing the displays in Fig. 7a illustrates that the distance between these fixed points for the two options increases with the confirmation bias, and this will be shown formally below. For a high level of confirmation bias illustrated in the right display of Fig. 7a, there exists another fixed point V_{false}^i for each value, corresponding to a false belief that the poorer option is superior. In the simulation illustrated in Fig. 7a, right, the estimated values initially fluctuate around V_{false}^i and in this period $V_t^- > V_t^+$. Due to stochastic nature of rewards the values may switch between fluctuations around V_{false}^i and V_{true}^i, and such shift happened around trial 500, in Fig. 7a, right. Importantly, we demonstrate formally below that these additional fixed points V_{false}^i only appear for the confirmation bias above a certain critical value, thus the confirmation model tends to get stuck in false belief only when the bias is higher than a specific value (dependent on task parameters).

We were not able to obtain tractable analytic expressions for stochastic fixed points of values when the softmax choice rule was assumed, hence we considered a simpler ε-greedy choice rule. We denote the probability of selecting an option with a lower estimated value by ε. To find the stochastic fixed points, we will assume that it rarely changes which of V_t^+ and V_t^- is higher. Indeed, in simulation of Fig. 7a, right, such change occurred only once in 1000 trials. Therefore, we will analyse the behaviour within the intervals on which $V_t^+ > V_t^-$, i.e. agent's beliefs on superiority of options are true, and within intervals on which $V_t^+ < V_t^-$, i.e. agent's beliefs are false.

Let us first consider a case of true beliefs, where a learned value for the richer option V_t^+ is higher than the value for the poorer option V_t^-. In this case, the richer option is selected with probability $1 - \varepsilon$, while the poorer option with probability ε.

The average change in the value of the richer option is then given by:

$$\begin{aligned}\mathrm{E}\left(\Delta V_t^+\right) = {} & (1-\varepsilon)\left[p^+\alpha^C\left(1-V_t^+\right) + (1-p^+)\alpha^D\left(-V_t^+\right)\right] \\ & + \varepsilon \quad \left[p^+\alpha^D\left(1-V_t^+\right) + (1-p^+)\alpha^C\left(-V_t^+\right)\right]\end{aligned} \tag{8}$$

In the above equation, the first line corresponds to changes occurring when the richer option is chosen, and the second line when the poorer option is chosen. Within each line, the first term in a square bracket corresponds to a change when the richer option yields rewards, while the second term when the richer option is not rewarded. To find the value in a stochastic fixed point, we set the left hand side of the above equation to 0 (because the stochastic fixed point is defined as the value in which the average value change is 0), and so the values in the fixed point V^{+}_{true} need to satisfy:

$$\begin{aligned} 0 = {} & (1-\varepsilon)\left[p^{+}\alpha^{C}\left(1-V^{+}_{true}\right)+(1-p^{+})\alpha^{D}\left(-V^{+}_{true}\right)\right] \\ & +\varepsilon \quad \left[p^{+}\alpha^{D}\left(1-V^{+}_{true}\right)+(1-p^{+})\alpha^{C}\left(-V^{+}_{true}\right)\right] \end{aligned} \tag{9}$$

Solving for V^{+}_{true} we obtain:

$$V^{+}_{true} = \frac{bp^{+}(1-\varepsilon)+p^{+}\varepsilon}{b(p^{+}(1-\varepsilon)+(1-p^{+})\varepsilon)+(1-p^{+})(1-\varepsilon)+p^{+}\varepsilon} \tag{10}$$

The above equation shows that the value V^{+}_{true} in a stochastic fixed point does not depend on the individual learning rates, but only on their ratio b, similarly as in a previous study (Caze & van der Meer, 2013). Analogous analysis shows that the stochastic fixed point for the poorer option is equal to:

$$V^{-}_{true} = \frac{bp^{-}\varepsilon+p^{-}(1-\varepsilon)}{b(p^{-}\varepsilon+(1-p^{-})(1-\varepsilon))+(1-p^{-})\varepsilon+p^{-}(1-\varepsilon)} \tag{11}$$

We now demonstrate that the confirmation bias increases V^{+}_{true} and decreases V^{-}_{true}. To evaluate the effect of increasing α^{C} relatively to α^{D} on V^{+}_{true}, we compute

$$\frac{dV^{+}_{true}}{db} = \frac{p^{+}(1-p^{+})(1-2\varepsilon)}{\left[b(p^{+}(1-\varepsilon)+(1-p^{+})\varepsilon)+(1-p^{+})(1-\varepsilon)+p^{+}\varepsilon\right]^{2}} \tag{12}$$

The above expression is non-negative because the denominator is non-negative (as it is a square) and the numerator is a product of non-negative terms. This derivative will be positive if $0 < p^{+} < 1$, and $\varepsilon < \frac{1}{2}$, i.e. when the rewards are non-deterministic, and the choice policy is not completely random. The derivative for V^{-}_{true} is equal to an analogous expression but with a negative sign:

$$\frac{dV^{-}_{true}}{db} = -\frac{p^{-}(1-p^{-})(1-2\varepsilon)}{\left[b(p^{-}\varepsilon+(1-p^{-})(1-\varepsilon))+(1-p^{-})\varepsilon+p^{-}(1-\varepsilon)\right]^{2}} \tag{13}$$

In summary, for stochastic rewards, the confirmation bias increases V^+_{true} and decreases V^-_{true}, and hence it magnifies the difference between these stochastic fixed points. This magnification of distance is visible in Fig. 7a where the gap between dark green and red lines increases across the displays.

Let us now consider the behaviour of the model under false beliefs, i.e. during the intervals when $V^+_t < V^-_t$. In this case, the poorer option is chosen on the majority of trials, because the agent falsely beliefs it has higher value. Furthermore, V^-_t is updated in the same way V^+_t was updated under the correct beliefs. Hence the fixed point under the false beliefs, V^-_{false}, is given by an expression analogous to that for V^+_{true} (Equation 10) but with p^+ replaced by p^-. Similarly, V^+_{false} is given by an expression analogous to that for V^-_{true} (Equation 11) but with p^- replaced by p^+. Consequently, V^-_{false} and V^+_{false} inherit from V^+_{true} and V^-_{true} their dependence on confirmation bias, i.e. V^-_{false} increases with the confirmation bias, while V^+_{false} decreases with the bias. Green and red curves in Fig. 7b plot the expressions for the stochastic fixed points for sample parameters. Without the confirmation bias ($b = 1$), the expressions for true and false fixed points coincide, and then diverge with confirmation bias.

Importantly, the fixed points based on false beliefs only exist when the agent has false beliefs. Thus the agent will tend to stay in these fixed points only if the false belief is satisfied in these fixed points, i.e. $V^+_{false} < V^-_{false}$. In Fig. 7b, this false belief is only satisfied to the right from the intersection of the bright curves, so the intersection occurs at a critical value of the confirmation bias in which $V^+_{false} = V^-_{false}$. The fixed points V^-_{false} and V^+_{false} only emerge for the confirmation bias above this critical value, and to highlight this, the curves plotting expressions for V^-_{false} and V^+_{false} are only shown in solid in Fig. 7b when they become fixed points.

The existence of fixed points V^-_{false} and V^+_{false} only above critical confirmation bias is confirmed in simulations shown in Fig. 7b. Blue and magenta curves show the mean estimated values at the end of simulations. The left display corresponds to simulations in which the values are initialized to a false belief. In this case the values stay in V^-_{false} and V^+_{false} for sufficiently high confirmation bias, but move to V^+_{true} and V^-_{true} for lower biases. The right display corresponds to a simulation in which the values are initialized to 0.5. In this case the values always move towards V^+_{true} and V^-_{true} for low bias, while for large bias, on some simulations they go to V^-_{false} and V^+_{false}, as indicated by larger error bars.

The critical value of bias in which the bifurcation occurs can be found by finding value of b_{crit} for which $V^+_{false} = V^-_{false}$. In general, an analytic expression for b_{crit} is excessively long and thus uninformative, but an insightful expression can be found for the special case of deterministic choices, i.e. $\varepsilon = 0$. In this case the stochastic fixed point become:

$$V^{-}_{false} = \frac{bp^{-}}{bp^{-} + (1 - p^{+})} \tag{14}$$

$$V^{+}_{false} = \frac{p^{+}}{b(1 - p^{+}) + p^{+}} \tag{15}$$

Equating Equations 14 and 15 and solving for bias we find the critical value of the confirmation bias:

$$b_{\text{crit}} = \sqrt{\frac{1}{\frac{p^{-}}{1} - 1}} \tag{16}$$

Inspecting the above equation, we observe that b_{crit} increases with p^{+} and decreases with p^{-}. Therefore, the larger the difference in reward probabilities of the two options, the higher the confirmation bias needs to be for the agent to get stuck in the false belief.

Effects of confirmation bias on reward rate

The analysis shown in Fig. 6 illustrates why the benefit of confirmation drops off as the bias tends to the extreme – it is because under extreme bias, the agent falls into a feedback loop whereby it confirms its false belief that the lower-valued bandit is in fact the best. Over multiple simulations, this radically increases the variance in performance and thus dampens overall average reward (Fig. 6c). However, it is noteworthy that this calculation is made under the assumption that all trials are made with equivalent response times. In the wild, incorrect choices may be less pernicious if they are made rapidly, if biological agents ultimately seek to optimise their reward per unit time (or reward rate).

Here, we relaxed this assumption and asked how the confirmatory bias affected overall reward *rates*, under the assumption that decisions are drawn to a close after a bounded accumulation process that is described by the drift-diffusion model. This allows us to model not only the choice probabilities but also reaction times.

Methods of simulations

We simulated a *reinforcement learning drift diffusion model* (RLDDM) in which the drift rate was proportional to the difference in value estimates between the two bandits (Pedersen, Frank, & Biele, 2017), which in turn depends on the update policy (confirmatory, disconfirmatory, or neutral). At each trial, the relative evidence x in favour of one of the two options is integrated over time, discretized in finite time step i, until it reaches a threshold a, implying the selection of the favoured option such that:

$$x_{i+1} = x_i + v_i * dt + c * \sqrt{dt} * \mathcal{N}(0, 1) \tag{17}$$

with x_0, the initial evidence defined as: $\chi_0 = \frac{a}{2}$, dt set to 0.001 and c to 0.1. The drift rate v_t, is linearly defined from the difference in values such that:

$$v_t = v_{\text{mod}} * \left(V_t^+ - V_t^-\right) \tag{18}$$

V_t^+ and V_t^- being the values at trial t, of the correct and incorrect options respectively. We used in our simulation a drift rate scaling parameter and a threshold values that make the drift-diffusion model to produce the same choice probabilities as the *softmax* policy with a temperature $\beta = 0.1$. In particular, the probability of making a correct choice by a diffusion model (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) is given by:

$$P_t^+ = \frac{1}{1 + \exp\left(-av_t/c^2\right)} \tag{19}$$

The above probability is equal to that in Equation 7 if $av_{mod}/c^2 = 1/\beta$. Thus, we set $v_{mod} = a = \sqrt{0.1}$. The values are updated exactly the same way as in the *confirmation* model (Equations 3 and 4). We employed the setting with 128 trials, used stable contingencies with reward probabilities equal to $p^- = 0.35$ and $p^+ = 0.65$.

Results of simulations

When we plotted the overall accuracy of the model, the results closely resemble those from previous analyses, as is to be expected (Fig. 8a). When we examined simulated reaction times, we observed that confirmatory learning leads to faster decisions (Fig. 8b). This follows naturally from the heightened difference in values estimates for each bandit, as shown in Fig. 6. Critically, however, responses were faster for both correct and incorrect trials. This meant that confirmatory biases have the potential to draw decisions to a more rapid close, so that unrewarded errors give way rapidly to new trials which have a chance of yielding reward. This was indeed the case: when we plotted reward rate as a function of confirmatory bias, there was a relative advantage over a neutral bias even for those more extreme confirmatory strategies that were detrimental in terms of accuracy alone (Fig. 8c). Thus, even a severe confirmatory bias can be beneficial to reward rates in the setting explored here. However, we note that this may be limited to the case explored here, where the ratio of reward to penalty is greater than one.

Comparison with alternative models

In order to clarify the role of the constraint imposed on the learning rates and of the counterfactual feedback, we performed simulations with three additional models, which differ from the confirmation model in the update of the values of the unchosen option. Table 1 compares how the learning rates depend on the choice and the sign of prediction error in the confirmation model and the alternative models.

All of the alternative models update the value estimate V_t^i of the chosen option similarly to the confirmation model, i.e. according to a delta rule with two learning rates: α^+ for positive

updates and α^- for negative updates. The three additional models differ in their updates of the value estimate of the unchosen option. The first model, referred to as the *Valence model*, updates the value estimate of the unchosen option with learning rates depending on the sign of prediction error analogously as for the chosen option. Thus in this model the learning rate only depends on the sign of prediction error but not on whether the option was chosen. The second model, referred to as the *Hybrid model*, updates the value of unchosen option using an unbiased learning rate defined as $\alpha^= = (\alpha^+ + \alpha^-)/2$. We refer to this model as hybrid, because the learning rate for unchosen option in this model is the average of those in the valence model and the confirmation model (with $\alpha^C = \alpha^+$ and $\alpha^D = \alpha^-$). The third model, referred to as *Partial feedback*, does not update the value of unchosen option, hence it can describe learning in tasks in which feedback is only provided for the chosen option. We define an agent with a positivity bias as one for whom $\alpha^+ > \alpha^-$, whereas an agent with a negativity bias has $\alpha^+ < \alpha^-$, and an agent with no bias (or a neutral setting) has $\alpha^+ = \alpha^-$.

Fig. 9 shows the performance of all four models separately for reward probabilities for both options $p^i < 0.5$ (left panels) and for reward probabilities for both options $p^i > 0.5$ (right panels). For low reward probabilities, the simulations gave very similar results in terms of favourable learning rates combinations (Fig. 9a, 9c, 9e and 9g) whereas it is not the case for high probabilities (Fig. 9b, 9d, 9f and 9h). Two analyses can be made based on this figure: first, Fig. 9a-f compare different models in the case of full feedback (for both the chosen and unchosen options), second Fig. 9g-h illustrate the case of partial feedback, and we analyse these two cases below.

In the case of full feedback, Fig. 9a-b show that confirmation bias in the confirmation model increases average reward irrespectively of the range of reward probabilities for the two options. The consistent effect of confirmation bias contrasts with the opposite effects of biases in learning rates in the valence model (Fig. 9c-d), where positivity bias is beneficial for low reward probabilities, while negativity bias is beneficial for high reward probabilities. These effects can be understood on the basis of a previous study mentioned in the Introduction (Caze & van der Meer, 2013). That study analysed reinforcement leaning model in which the learning rate depended on the sign of prediction error as in the valence model. The study showed that if reward probabilities for both options $p^i < 0.5$, then it is beneficial to have a positivity bias. With such bias, both V^1 and V^2 will be overestimated and critically the difference $V^1 - V^2$ will be magnified, so with a noisy choice rule the option with the higher reward probability will be more likely to be selected. By contrast, if both $p^i > 0.5$, then overestimating V^1 and V^2 would actually reduce the difference $V^1 - V^2$ due to a ceiling effect, because according to Eq 1 reward estimates cannot exceed the maximum reward available, i.e. $V^i \leq 1$. In this case, it is beneficial to have a negativity bias. Therefore, if one assumes that learning rates can differ between rewarded and unrewarded trials, the type of reward-increasing bias depends on magnitude of reward probabilities in a task (Caze & van der Meer, 2013) and this dependence is clearly seen in Fig. 9c-d.

Since the learning rates in the hybrid model lie in between those in the valence and the confirmation models, the optimal bias in the hybrid model is in between that in these two models. Namely the positivity or confirmation bias is optimal for low reward probabilities

(Fig. 9e), while for high reward probabilities the optimal bias is close to $b \approx 1$ (Fig. 9f), so it is between the optimal biases for the confirmation (Fig. 9b) and valence (Fig. 9d) models.

It is also worth comparing the performance of the models for their optimal learning rates. Different panels in Fig. 9a-f have different colour scales which span the range of obtained rewards. Comparing colour scales reveals that the confirmation model can produce overall highest reward, namely for low reward probabilities it achieves higher reward (for its best parameters) than the valence model, and similar performance to the hybrid model, while for high reward probabilities, it obtained higher reward than both alternative models.

In summary, in the case of full feedback, the *Confirmation model* is the only one among models compared in Fig. 9a-f for which the optimal learning rates lie in the same regions of parameter space for both low and high probabilities. A learner often does not know the task parameters, and the confirmation model is most robust to this uncertainty, because it is the only model for which it is possible to choose a combination of learning rates that work relatively well for different tasks.

In the case of partial feedback where only the value of the chosen option is modified, the positivity bias is beneficial for low reward probabilities, while the negativity bias is beneficial for high reward probabilities (Fig. 9g-h), as expected from a previous theoretic analysis (Caze & van der Meer, 2013). The optimal learning rates with partial feedback are similar to those in the valence model with full feedback (cf. Fig. 9c-d and Fig. 9g-h) as in both models the learning rate only depends on the sign of prediction error (Table 1).

The optimal learning rates slightly differ between the valence model with full feedback and partial feedback, i.e., less negativity bias is required to maximize reward with partial feedback for high probabilities (c.f. Fig. 9d and 9h). This difference arises because with full feedback both values are updated equally often, while with partial feedback the poorer option is chosen less frequently. Hence with partial feedback the value of the poorer option moves slowly from its initial value of 0.5, so even if $\alpha^+ > \alpha^-$, the value of the poorer option may not be overestimated. The difference between the models disappears if both values are updated with more similar frequencies (we observed it in simulations (not shown) in which temperature of the softmax function was increased).

In summary, in the case of partial feedback, updating values of the chosen option with the larger learning rate after positive prediction error is detrimental for higher reward probabilities (Fig. 9h). Hence the bias which optimizes the confirmation model (Fig. 9b) may be detrimental with partial feedback in the model analysed in this section (Fig. 9h). Nevertheless, in the Discussion we will come back to this issue, and point out that the optimal bias may differ in other reinforcement learning models with partial feedback.

Discussion

Humans have been observed to exhibit confirmatory biases when choosing between stimuli or actions that payout with uncertain probability (Chambon et al., 2020; Palminteri et al., 2017; Schuller et al., 2020). These biases drive participants to update positive outcomes (or those that are better than expected) for chosen options more sharply than negative

outcomes, but to reverse this update pattern for the unchosen option. Here, we show through simulations that in an extended range of settings traditionally used in human experiments, this asymmetric update is advantageous in the presence of noise in the decision process. Indeed, agents who exhibited a confirmatory bias, rather than a neutral or disconfirmatory bias, were in most circumstances tested those agents that reaped the largest quantities of reward. This counterintuitive result directly stems from the update process itself that biases the value of the chosen and unchosen options (corresponding overall to the best and worst options respectively), increasing mechanistically their relative distance from each other and ultimately the probability of selecting the best option in the upcoming trials.

Exploring the evolution of action values under confirmatory updates offers an insight into why this occurs. Confirmatory updating has the effect of rendering subjective action values more extreme than their objective counterparts – in other words, options that are estimated to be good are overvalued, and options estimated to be bad are undervalued (**Fig. 6**). This can have both positive and negative effects. The negative effect is that a sufficiently strong confirmatory bias can drive a feedback loop whereby poor or mediocre items that are chosen by chance can be falsely updated in a positive direction, leading to them being chosen more often. The positive effect, however, is that where decisions are themselves intrinsically variable (for example, because they are corrupted by Gaussian noise arising during decision-making or motor planning, modelled here with the softmax temperature parameter) overestimation of value makes decisions more robust to decision noise, because random fluctuations in the value estimate at the time of the decision are less likely to reverse a decision away from the better of the two options. The relative strength of these two effects depends on the level of decision noise: within reasonable noise ranges the latter effect outweighs the former and performance benefits overall.

Relationship to other studies

The results described here thus join a family of recent-reported phenomena whereby decisions that distort or discard information lead to reward-maximising choices under the assumption that decisions are made with finite computational precision – i.e. that decisions are intrinsically noisy (Summerfield & Tsetsos, 2015). For example, when averaging features from a multi-element array to make a category judgment, under the assumption that features are equally diagnostic (and that the decision policy is not itself noisy), then normatively, they should be weighted equally in the choice. However, in the presence of "late" noise, encoding models that overestimate the decision value of elements near the category boundary are reward-maximising, for the same reason as the confirmatory bias here: they inflate the value of ambiguous items away from indifference, and render them robust to noise (Li, Herce Castanon, Solomon, Vandormael, & Summerfield, 2017). A similar phenomenon occurs when comparing gambles defined by different monetary values: utility functions that inflate small values away from indifference (rendering the subjective difference between $2 and $4 greater than the subjective difference between $102 and $104) have a protective effect against decision noise, providing a normative justification for convex utility functions (Juechems, Spitzer, Balaguer, & Summerfield, 2020). Related results have been described in problems that involve sequential sampling in time, where they may account for violations of axiomatic rationality, such as systematically intransitive

choices (Tsetsos et al., 2016). Moreover, a bias in how evidence is accumulated within a trial has been shown to increase the accuracy of individual decisions, making the decision variable more extreme and thus less likely to be corrupted by noise (Zhang & Bogacz, 2010).

Recent studies also report simulations of the confirmation bias model (Chambon et al., 2020; Tarantola, Folke, Boldt, Perez, & De Martino, 2021). These simulations paralleled experimental paradigms reported in these papers, and a confirmation model was simulated for parameters (including softmax temperature) corresponding to those estimated for participants of the studies. The simulated agents employing confirmation bias obtained higher average reward than unbiased learners, as well as learners described by other models. Our paper suggests the same conclusion using a complementary approach in which the models have been simulated in variety of conditions and analysed mathematically.

Modelling studies have investigated how learning with rates depending on the sign of prediction error could be implemented in the basal ganglia circuits known to underlie reinforcement learning (Collins & Frank, 2014; Dabney et al., 2020). Models have been developed to describe how positive and negative prediction errors preferentially engage learning in different populations of striatal neurons (Mikhael & Bogacz, 2016; Möller & Bogacz, 2019). It would be interesting to investigate the neural mechanisms that lead to learning rates depending not only on the sign of prediction error but also on whether options have been chosen.

Validity of model's assumptions

Reinforcement learning models fit to human data often assume that choices are stochastic, i.e. that participants fail to choose the most valuable bandit. In standard tasks involving only feedback about the value of the chosen option (factual feedback), some randomness in choices promotes exploration which in turns allows information to be acquired that may be relevant for future decisions. However, our task involves both factual and counterfactual feedback, and so exploration is not required to learn the value of the two bandits. Nevertheless, in some simulations we modelled choices with a softmax rule, which assumes that decisions are corrupted by Gaussian noise, or an ε-greedy policy, which introduces lapses to the choice process with a fixed probability. Implicitly, thus, we are committing to the idea that value-guided decisions may be irreducibly noisy even where exploration is not required (Renart & Machens, 2014). Indeed, others have shown that participants continue to make noisy decisions even where counterfactual feedback is available, even if they have attributed that noise to variability in learning rather than choice (Findling, Skvortsova, Dromnelle, Palminteri, & Wyart, 2019).

Due to our assumptions, the present study has a number of limitations. Firstly, we explored the properties of a confirmatory model that has been previously shown to provide a good fit to human data performing a bandit task with factual and counterfactual feedback. However, we acknowledge that this is not the only possible model that could increase reward by enhancing the difference between represented values of options. In principle, any other models producing choice hysteresis might be able to explain these results (Katahira, 2018; Miller, Shenhav, & Ludvig, 2019; Worthy, Pang, & Byrne, 2013). An analysis of these

Europe PMC Funders Author Manuscripts

different models and of their respective resilience to decision noise in different settings is beyond the scope of the current study but would be an interesting target for future research. Secondly, the results described here hold assuming a fixed and equal level of stochasticity (e.g. *softmax* temperature) in agents' behaviours, irrespective of their bias (i.e. the specific combination of learning rates). Relaxing this assumption, an unbiased agent could perform equally well as a biased agent subject to more decision noise. Thus, the benefit of confirmatory learning is relentlessly linked to the level of noise and one level of confirmation bias cannot be thought as being beneficial overall. Thirdly, the present study does not investigate the impact on the performance of other kinds of internal noise such like an update noise (Findling et al., 2019). The latter, instead of perturbing the policy itself, perturbs at each trial the update process of the options' value (i.e. predictions errors are blurred with a Gaussian noise), and cannot presumably produce a similar increase in performance, having overall no effect on the average difference between these option values.

Confirmation bias with partial feedback

In this paper we focused on studying confirmation bias in tasks where the feedback is provided for both chosen and unchosen options, but in most reinforcement learning tasks studied in the laboratory and possibly in the real world, feedback is provided only for the chosen option. With such partial feedback it seems not possible to distinguish between the confirmation and valence models because they make the same update of the value of the chosen option. However, a recent ingenious experiment suggested that the confirmation bias was also present with partial feedback, because the learning rate was higher after positive prediction errors only if the choice was made by the participant, but not when the choice was made by a computer (Chambon et al., 2020). Analogous effect was also observed outside the laboratory in a study of heart surgeons, who learned more from their own successes than failures, but not from observed successes of their colleagues (Kc, Staats, & Gino, 2013). Hence it is important to understand how results from this paper could be generalized to partial feedback.

For partial feedback, previous theoretic work suggests that optimal leaning rates depend on whether the reward probability is high or low (Caze & van der Meer, 2013) and we confirmed it in simulations in Fig. 9g-h. Surprisingly, it has been shown that human participants did not follow this pattern, and had similar learning rates irrespectively if reward probabilities were low or high (Chambon et al., 2020; Gershman, 2015). This poses a question whether humans do not behave in a way maximizing rewards (which seems unlikely given the evolutionary pressure for reward maximization), or the normative theory of learning with partial feedback needs to be revised. One way to include confirmation bias in models of learning with partial feedback would be to note that humans and animals are aware of confidence of their choices (Kiani & Shadlen, 2009), i.e. whether they are certain the chosen option yields highest reward or if the choice was a guess. Hence one could consider models in which learning rate depends not only the sign of prediction error but also on the confidence, such that the negative feedback is taken into consideration less when a participant is confident of their choice. Formulating such models would require careful comparison of the models with specially designed experiments, hence it is beyond the scope of this paper, but would be an interesting direction for future work.

Limits to the benefits of biased beliefs

It is important to point out that confirmation bias is beneficial in many but not all circumstances. Although in almost all presented simulations there exists a combination of biased learning rates giving performance that is higher or as good as the best unbiased learner, the optimal learning rates and hence the amount of bias differ depending on task parameters. At the start of the task, a learner usually is unable to know the details of the task a priori, so needs to adopt a certain default combination of learning rates. One could expect that such default learning rates would be determined by past experience or even be to a certain extent influenced by evolution. However such default set of biased learning rates will lead to detrimental effects on performance in certain tasks. For example, a recent study estimated average learning rates of human participants to be $\alpha^C \approx 0.15$ and $\alpha^D \approx 0.05$ (Tarantola et al., 2021) giving a confirmation bias of $b \approx 3$. Although such strong confirmation bias increases reward in many simulated scenarios when decisions are noisy (e.g. Fig. 3e-h), it would have a negative effect on performance when decisions are accurately made on the basis of values and in changing environments (e.g. Fig. 3a-d). If the default confirmation bias is influenced by evolution, its value is likely to be relatively high, because many of the key decisions of our ancestors had to be quick and thus were noisy due to the speed accuracy trade-off. By contrast, in the modern world, we often can take time to consider important choices, hence the biases that brought evolutionary advantage to our ancestor may not always be beneficial to modern humans.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work has been supported by MRC grants MC_UU_12024/5, MC_UU_00003/1, BBSRC grant BB/S006338/1, and ERC Consolidator grant 725937.

References

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev. 2006; 113 (4) 700–765. [PubMed: 17014301]

Caze RD, van der Meer MA. Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern. 2013; 107 (6) 711–719. DOI: 10.1007/s00422-013-0571-5 [PubMed: 24085507]

Chambon V, Théro H, Vidal M, Vandendriessche H, Haggard P, Palminteri S. Information about action outcomes differentially affects learning from self-determined versus imposed choices. Nature Human Behaviour. 2020; 4 (10) 1067–1079.

Cieślak PE, Ahn W-Y, Bogacz R, Parkitna JR. Selective effects of the loss of NMDA or mGluR5 receptors in the reward system on adaptive decision-making. Eneuro. 2018; 5 (4)

Collins AG, Frank MJ. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychological review. 2014; 121 (3) 337. [PubMed: 25090423]

Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, Botvinick M. A distributional code for value in dopamine-based reinforcement learning. Nature. 2020; 577 (7792) 671–675. [PubMed: 31942076]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441 (7095) 876–879. [PubMed: 16778890]

Farashahi S, Donahue CH, Hayden BY, Lee D, Soltani A. Flexible combination of reward information across primates. Nature Human Behaviour. 2019; 3 (11) 1215–1224.

Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V. Computational noise in reward-guided learning drives behavioral variability in volatile environments. Nat Neurosci. 2019; 22 (12) 2066–2077. DOI: 10.1038/s41593-019-0518-9 [PubMed: 31659343]

Gershman SJ. Do learning rates adapt to the distribution of rewards? Psychon Bull Rev. 2015; 22 (5) 1320–1327. DOI: 10.3758/s13423-014-0790-3 [PubMed: 25582684]

Groopman, J. How Doctors Think. Mariner Books; 2007.

Juechems K, Spitzer B, Balaguer J, Summerfield C. Optimal utility and probability functions for agents with finite computational precision. PsyArXiv. 2020.

Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. J Math Psychol. 2018; doi: 10.1016/j.jmp.2018.09.002

Kc D, Staats BR, Gino F. Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. Management Science. 2013; 59 (11) 2435–2449.

Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. Science. 2009; 324 (5928) 759–764. [PubMed: 19423820]

Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S. Behavioural and neural characterization of optimistic reinforcement learning. Nat Hum Behav. 2017; 1

Li V, Herce Castanon S, Solomon JA, Vandormael H, Summerfield C. Robust averaging protects decisions from noise in neural computations. PLoS Comput Biol. 2017; 13 (8) e1005723 doi: 10.1371/journal.pcbi.1005723 [PubMed: 28841644]

Mikhael JG, Bogacz R. Learning reward uncertainty in the basal ganglia. PLoS computational biology. 2016; 12 (9) e1005062 [PubMed: 27589489]

Miller KJ, Shenhav A, Ludvig EA. Habits without values. Psychol Rev. 2019; 126 (2) 292–311. DOI: 10.1037/rev0000120 [PubMed: 30676040]

Möller M, Bogacz R. Learning the payoffs and costs of actions. PLoS computational biology. 2019; 15 (2) e1006285 [PubMed: 30818357]

Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Review of General Psychology. 1998; 2: 175–220.

Niv Y, Edlund JA, Dayan P, O'Doherty JP. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci. 2012; 32 (2) 551–562. DOI: 10.1523/JNEUROSCI.5498-10.2012 [PubMed: 22238090]

Oaksford M, Chater N. Optimal data selection: revision, review, and reevaluation. Psychon Bull Rev. 2003; 10 (2) 289–318. DOI: 10.3758/bf03196492 [PubMed: 12921410]

Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. PLoS Comput Biol. 2017; 13 (8) e1005684 doi: 10.1371/journal.pcbi.1005684 [PubMed: 28800597]

Pedersen ML, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. Psychon Bull Rev. 2017; 24 (4) 1234–1251. DOI: 10.3758/s13423-016-1199-y [PubMed: 27966103]

Renart A, Machens CK. Variability in neural activity and behavior. Curr Opin Neurobiol. 2014; 25: 211–220. DOI: 10.1016/j.conb.2014.02.013 [PubMed: 24632334]

Rescorla, RA, Wagner, AR. Classical Conditioning II: Current Research and Theory. AH, B, Prokasy, WF, editors. Appleton Century Crofts; New York: 1972. 64–99.

Schuller T, Fischer AG, Gruendler TOJ, Baldermann JC, Huys D, Ullsperger M, Kuhn J. Decreased transfer of value to action in Tourette syndrome. Cortex. 2020; 126: 39–48. DOI: 10.1016/j.cortex.2019.12.027 [PubMed: 32062469]

Summerfield C, Tsetsos K. Do humans make good decisions? Trends Cogn Sci. 2015; 19 (1) 27–34. DOI: 10.1016/j.tics.2014.11.005 [PubMed: 25488076]

Talluri BC, Urai AE, Tsetsos K, Usher M, Donner TH. Confirmation bias through selective overweighting of choice-consistent evidence. Current Biology. 2018; 28 (19) 3128–3135. e3128 [PubMed: 30220502]

Tarantola TO, Folke T, Boldt A, Perez OD, De Martino B. Confirmation bias optimizes reward learning. bioRxiv. 2021.

Tsetsos K, Moran R, Moreland J, Chater N, Usher M, Summerfield C. Economic irrationality is optimal during noisy decision making. Proc Natl Acad Sci U S A. 2016; 113 (11) 3102–3107. DOI: 10.1073/pnas.1519157113 [PubMed: 26929353]

Worthy DA, Pang B, Byrne KA. Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. Front Psychol. 2013; 4: 640. doi: 10.3389/fpsyg.2013.00640 [PubMed: 24137137]

Zhang J, Bogacz R. Bounded Ornstein–Uhlenbeck models for two-choice time controlled tasks. Journal of Mathematical Psychology. 2010; 54 (3) 322–333. DOI: 10.1016/j.jmp.2010.03.001

Europe PMC Funders Author Manuscripts

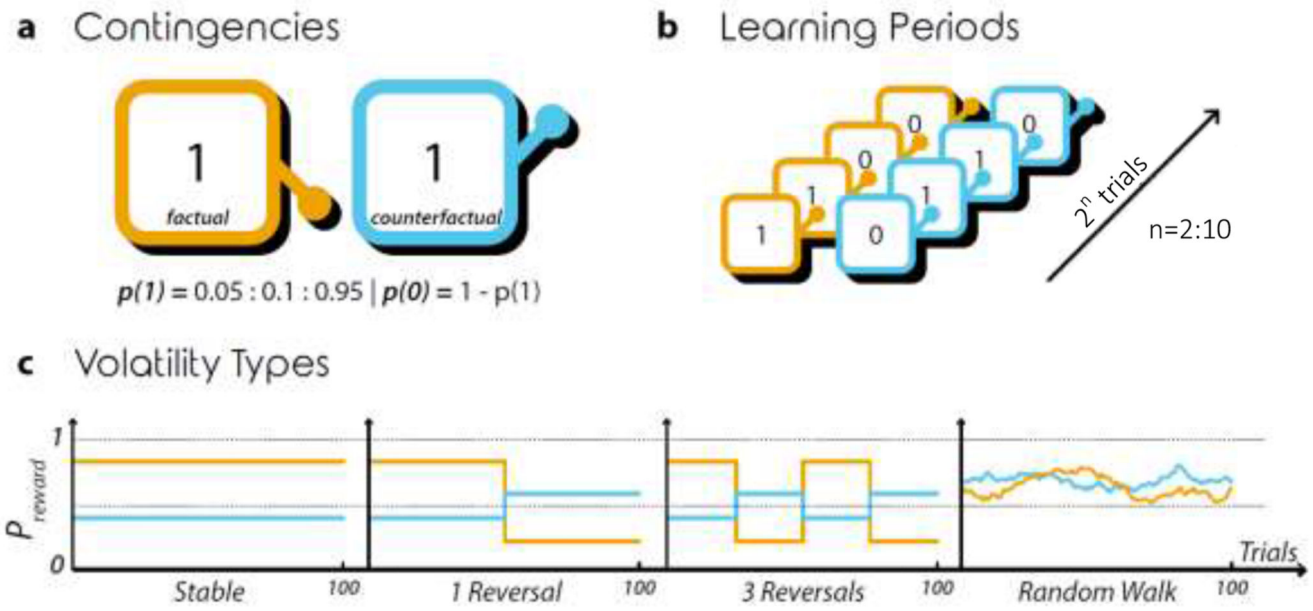

Figure 1. Simulation Setup.

a. Reward contingencies. The illustration represents the chosen (orange) and unchosen (blue) bandits each with a feedback signal (central number). Below, we state the range of possible outcomes and probabilities. **b**. Learning Periods. The illustration represents the different length of the learning period and the different outcomes combinations potentially received by the agents. **c**. Volatility Types. The line plots represent the evolution of the two arms probability across trials in the different volatility conditions.

Europe PMC Funders Author Manuscripts

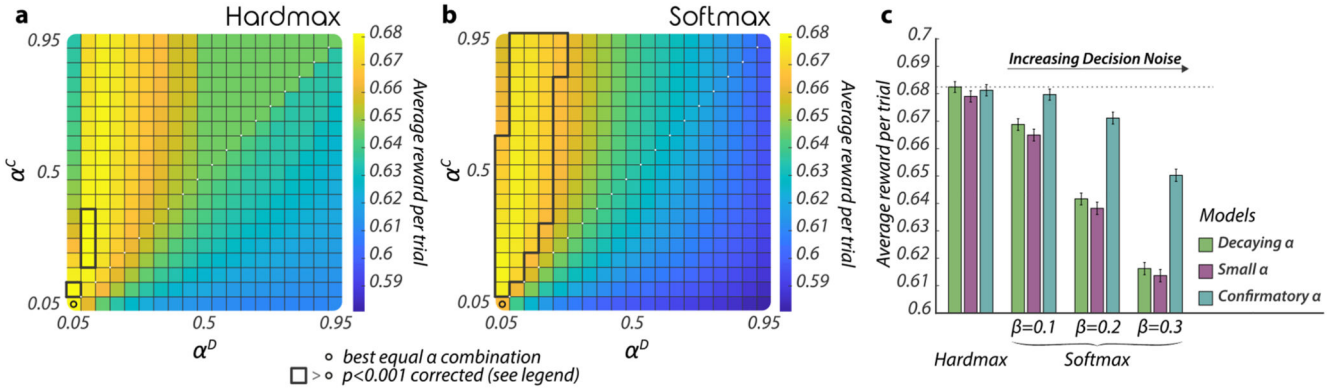

Figure 2. Dependence of reward on learning rate and decision noise in a stable environment.

a and **b**. Average reward for all learning rate combinations. The heatmaps represent the per trial average reward for combinations of α^C (y-axis) and α^D (x-axis), averaged across all reward contingencies and agents in the stable condition with 1024 trials. Areas enclosed by black lines represent learning rate combinations for which the reward is significantly higher than the performance of the best equal learning rates combination represented by a black circle, one-tailed independent samples rank-sum tests, p<0.001 corrected for multiple comparison. **a**. Deterministic Decisions. Simulated reward is obtained using a noiseless *hardmax* policy. **b**. Noisy Decisions. Simulated reward is obtained using a noisy *softmax* policy with $\beta = 0.1$. **c**. Comparison with optimal models. The bar plot represents the per trial average reward of the *confirmation* model, the *small learning rate* model and the *decaying learning rate* model for four different levels of noise in the decision process. In simulations of the *confirmation model*, the best learning rates combination was used for each noise level (i.e. α^C = [0.1,0.15, 0.3,0.35] and α^D = 0.05). Bars represent the means and error bars the standard deviations across agents, all reward levels are significantly different from each other, two-tailed independent samples rank-sum tests, p<0.001.

Europe PMC Funders Author Manuscripts

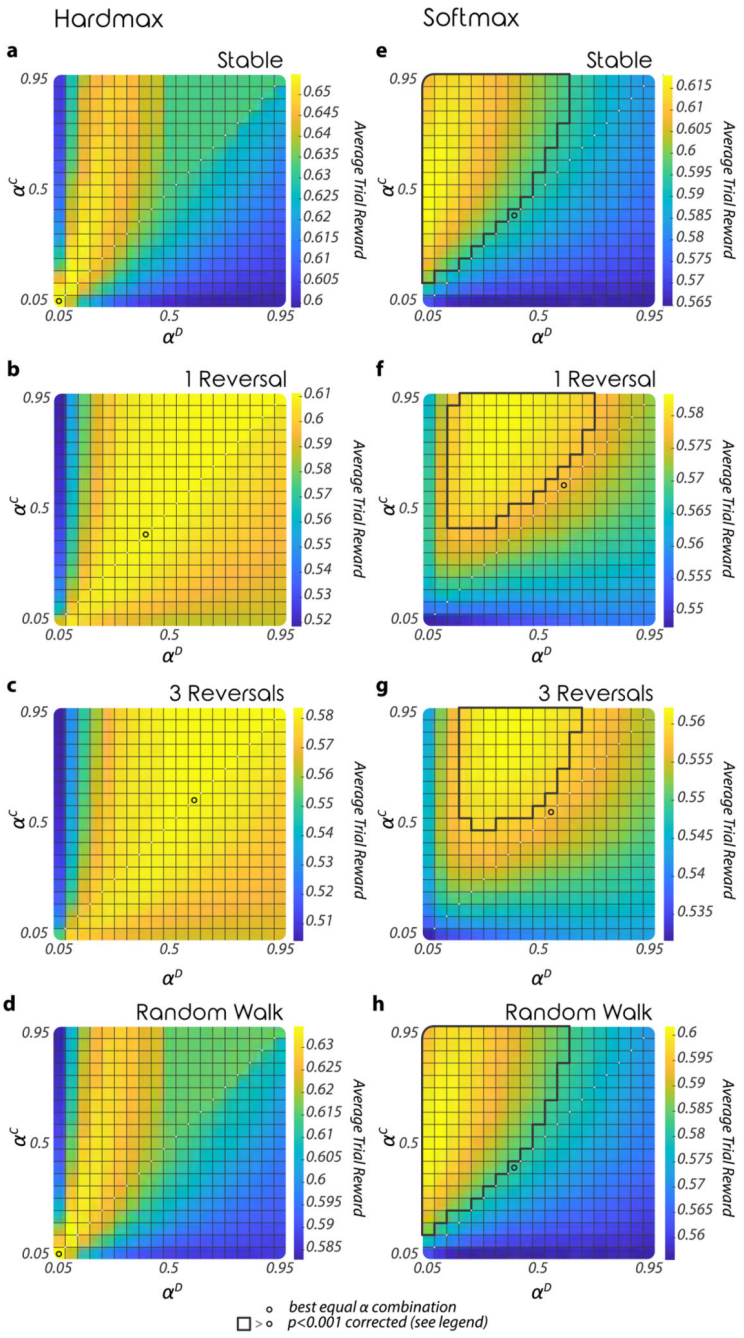

Figure 3. Dependence of reward on learning rate and decision noise in different environments.

a, b, c, d, e, f, g & h. The heatmaps represent the per trial average reward for combinations of α^C (y-axis) and α^D (x-axis) given a *hardmax* policy (**a, b, c, d**) or a *softmax* policy ($\beta = 0.3$) (**e, f, g, h**). The performance is averaged across all reward contingencies, period lengths and 1000 agents in the stable condition (**a**, **e**), 1 reversal condition (**b**, **f**), 3 reversals condition (**c**, **g**) or 100000 agents in the random walk condition (**d, h**). Areas enclosed by black lines represent learning rate combinations for which the reward is significantly higher than the reward of the best equal learning rates combination represented by a black

circle, one-tailed independent samples rank-sum tests, $p < 0.001$ corrected for multiple comparisons.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

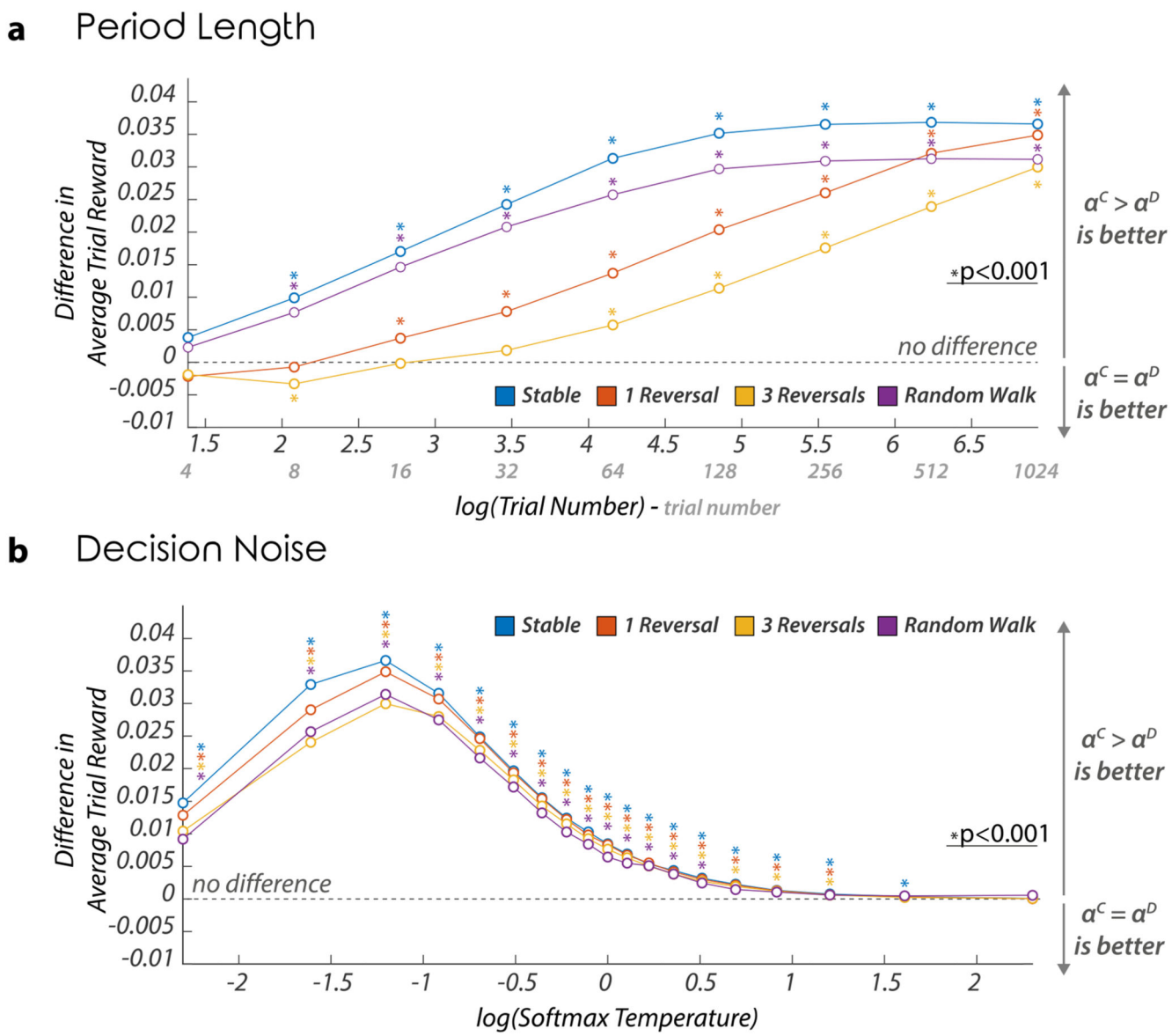

Figure 4. Effects of period length and decision noise on the relative performance of confirmation model.

a. Effect of period length on reward. The line plot represents the difference in average reward between the confirmation model (with the best confirmatory learning rate combination per period) and the unbiased model (with the best per period single learning rate) in function of the log of the period length, and for the four different volatility conditions. The logarithmic transformation of the trial number is for illustrative purpose only. *, $p < 0.001$, two-tailed independent rank-sum tests. **b**. Effect of decision noise on performance. The line plot represents the difference in per trial average performances of the confirmation model (with the best confirmatory learning ratescombination) and the unbiased model (with the single best learning rate) in function of the log of *softmax* temperature, and for the four different volatility conditions. The logarithmic transformation of the *softmax*

temperature is for illustrative purpose only. *, $p < 0.001$, two-tailed independent rank-sum tests.

Europe PMC Funders Author Manuscripts

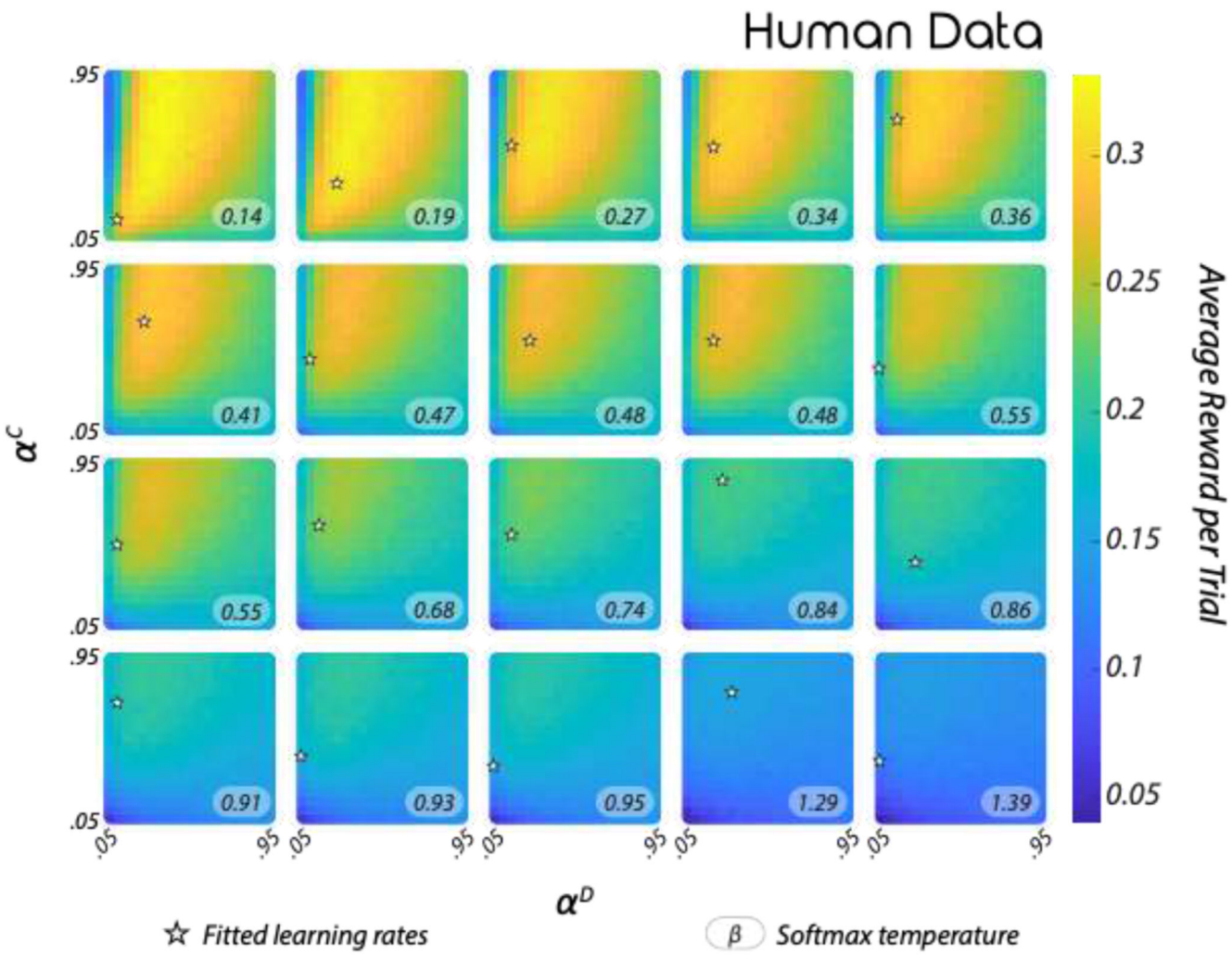

Figure 5. Relation between human and synthetic data.

The heatmaps represent the per trial average reward for combinations of α^C (y-axis) and α^D (x-axis) in the experimental environment studied by Palminteri et al. (2017). Simulations have been performed with different *softmax* temperatures corresponding to the fitted temperature of the participants from that study and are averaged across 1000 agents. The stars represent for each participant, the combination of fitted learning rates.

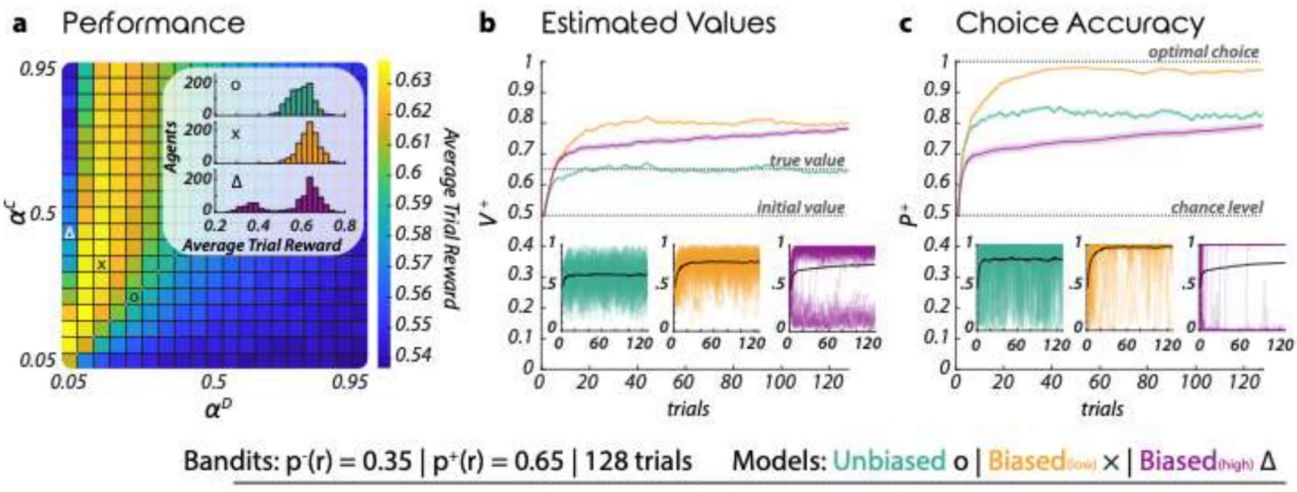

Figure 6. Mechanism by which confirmation bias tends to increase reward.

a. Average reward and reward distributions for different levels of confirmation bias. The heatmap represents the per trial average reward of the confirmation model for all learning rates combinations (confirmatory learning rates are represented on the y-axis whereas disconfirmatory learning rates are represented on the x-axis) associated with a *softmax* policy with $\beta = 0.1$. The rewards concern the stable condition with 128 trials and asymmetric contingencies ($p^- = 0.35$ and $p^+ = 0.65$) and are averaged across agents. The three signs inside the heatmap (Δ, × and +) represent the three learning rates combinations used in the simulations illustrated in panels **b** and **c**. The histograms show the distribution across agents of the average per trial reward for the three different combinations. **b**. Estimated values. The line plots represent the evolution of the best option value V^+ across trials. The large plot represents the agents-averaged value of the best option across trials for three different learning rates combinations, “unbiased” ($a^C = a^D = 0.25$), “biased (low)” ($a^C = 0.35$ and $a^D = 0.15$) and “biased (high)” ($a^C = 0.45$ and $a^D = 0.05$). The lines represent the mean and the shaded areas, the SEM. The small plots represent the value of the best option across trials plotted separately for the three combinations. The thick lines represent the average across agents and the lighter lines the individual values of 5% of the agents. **c**. Choice Accuracy. The line plots represent the evolution of the probability to select the best option across trials. The large plot represents the agents-averaged probability to select the best option across trials for three different learning rates combinations, “unbiased” ($a^C = a^D = 0.25$), “biased (low)” ($a^C = 0.35$ and $a^D = 0.15$) and “biased (high)” ($a^C = 0.45$ and $a^D = 0.05$). The lines represent the mean and the shaded areas, the SEM. The small plots represent the probability to select the best option across trials plotted separately for the three combinations. The thick lines represent the average across agents and the lighter lines the individual probability for 5% of the agents.

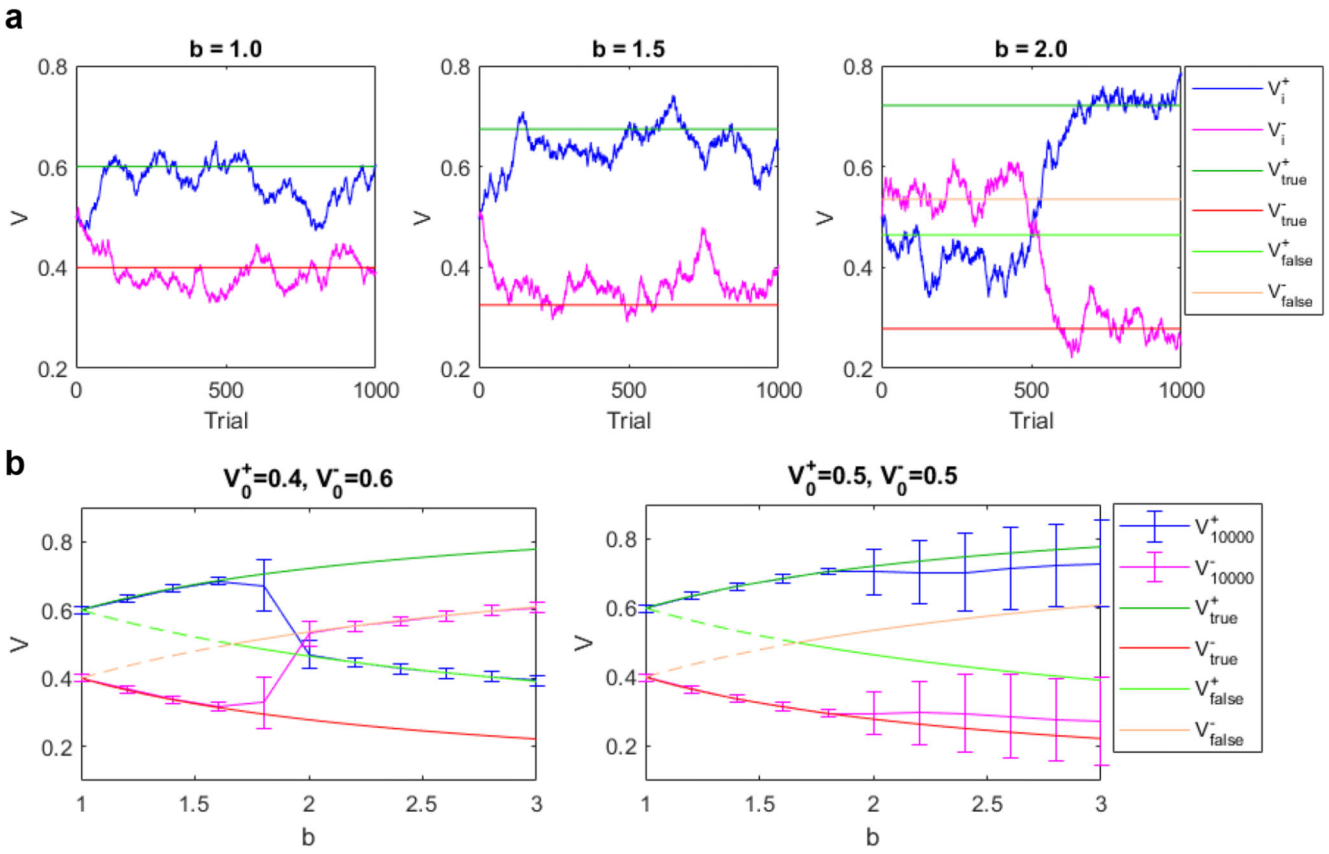

Figure 7. Stochastic fixed points of value estimates.

Behaviour of the confirmation model with ε-greedy choice policy ($\varepsilon = 0.1$) has been analysed for a stable environment with reward probabilities of the two options equal to $p^+ = 0.6$ and $p^- = 0.4$. **a**. Blue and purple lines show the evolution of value estimates over simulated trials. Different displays correspond to different level of confirmation bias b, indicated above the displays. The learning rates were set to $\alpha^D = 0.01$ and $\alpha^C = b\alpha^D$. **b**. Asymptotic behaviour of the confirmation model for different level of the confirmation bias. The blue and magenta curves show the average estimated values at the end simulation with 10,000 trials. This average is taken over 100 simulations, and the error bars indicate the standard deviation. The model was simulated with $\alpha^D = 0.01$ and $\alpha^C = b\alpha^D$, where the confirmation bias b is shown on x-axes. Red and green curves denote the values of stochastic fixed points. The two displays correspond to different initial estimated values, listed above the displays.

Europe PMC Funders Author Manuscripts

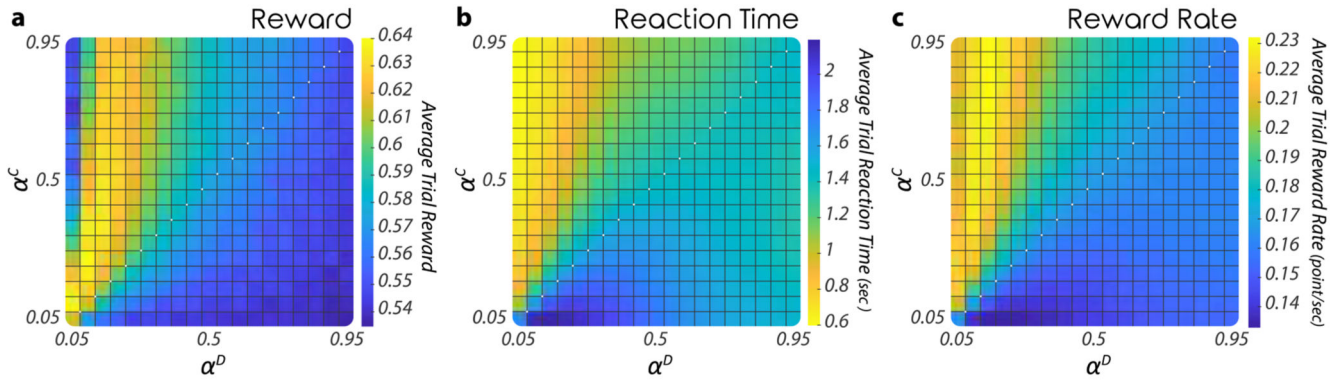

Figure 8. Effect of confirmation bias on reward rate.

a. The heatmap represents the per trial average reward simulated with the confirmation RLDDM for all learning rates combinations (confirmatory learning rates are represented on the y-axis whereas disconfirmation learning rates are represented on the x-axis). The rewards concern the stable condition with 128 trials and asymmetric contingencies ($p^- = 0.35$ and $p^+ = 0.65$) and are averaged across agents. **b**. The heatmap represents the per trial average reaction time estimated with the confirmation RLDDM for all learning rates combinations. **c**. The heatmap represents the per trial average reward rate simulated with the confirmation RLDDM for all learning rates combinations.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

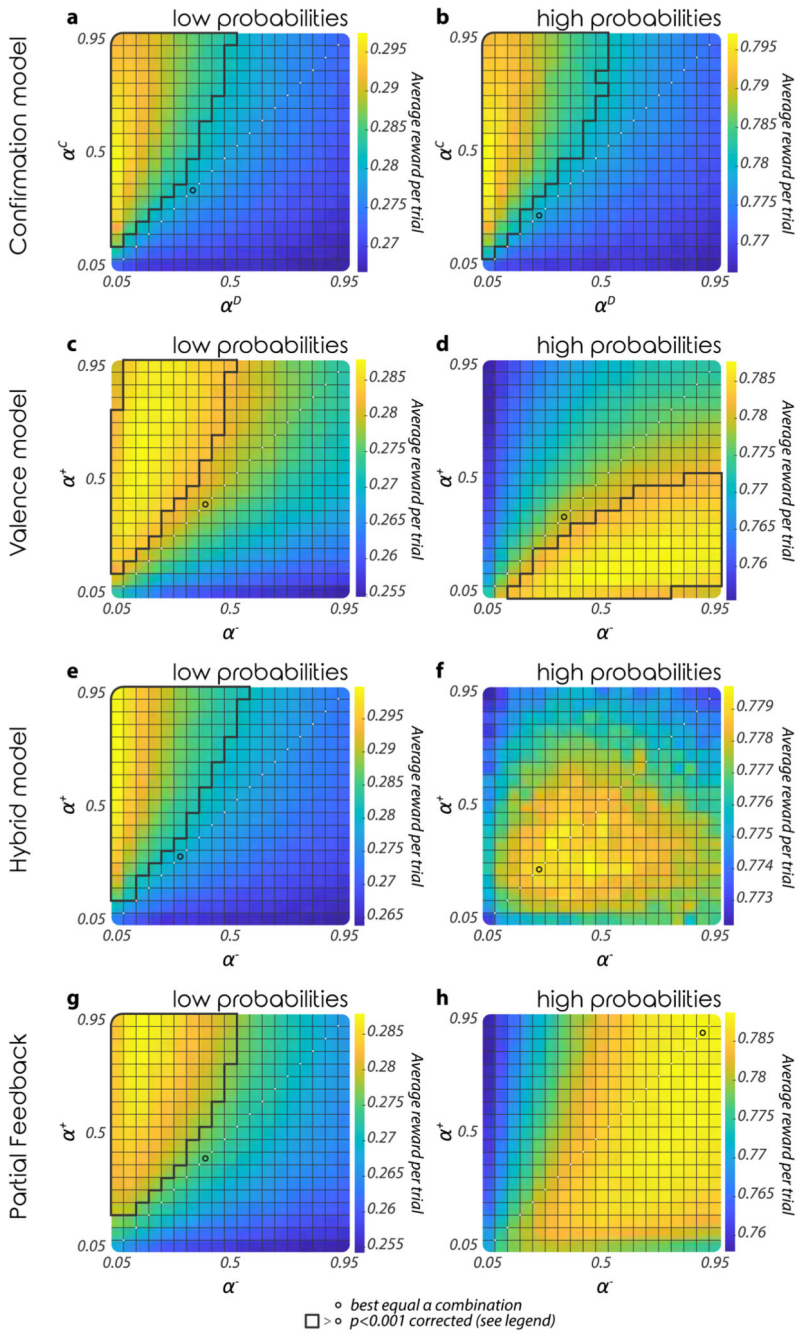

Figure 9. Dependence of reward on contingencies in alternative models.

a - h. The heatmaps represent the per trial average reward for combinations of α^C (y-axis) and α^D (x-axis) (**a & b**) or α^+ (y-axis) and α^- (x-axis) (**c - h**) with a *softmax* policy (β = 0.3) in a stable environment. The performance is averaged across 1000 agents, all period lengths and low reward contingencies, i.e. $p^1 < 0.5$ and $p^2 < 0.5$ (**a**, **c**, **e** & **g**) or high reward contingencies, i.e. $p^1 > 0.5$ and $p^2 > 0.5$ (**b, d, f & h**). The four models are the *Confirmation* model (**a & b**), the *Valence* model (**c & d**) *Hybrid* model (**e & f**) and a model with partial feedback (**g & h**). Areas enclosed by black lines represent learning rate combinations for

which the reward is significantly higher than the reward of the best equal learning rates combination represented by a black circle, one-tailed independent samples rank-sum tests, $p < 0.001$ corrected for multiple comparisons.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Learning rates in the confirmation and alternative models. To make the table easier to read, α^C and α^D are highlighted in bold.

Model	Chosen option i		Unchosen option $j \neq i$	
	$\delta_t^i > 0$	$\delta_t^i < 0$	$\delta_t^j > 0$	$\delta_t^j < 0$
Confirmation model	$\boldsymbol{\alpha^C}$	α^D	α^D	$\boldsymbol{\alpha^C}$
Valence model	$\boldsymbol{\alpha^+}$	α^-	$\boldsymbol{\alpha^+}$	α^-
Hybrid model	$\boldsymbol{\alpha^+}$	α^-	$\alpha^=$	$\alpha^=$
Partial feedback	$\boldsymbol{\alpha^+}$	α^-	-	-