A normative account of confirmatory biases during reinforcement learning

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

[1] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK.

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

[†] Equal contribution

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

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 overestimate the value of more valuable bandits, and underestimate the value of less valuable bandits, rendering decisions overall more robust in the face of noise. Our results show how apparently suboptimal learning policies can in fact be reward-maximising if decisions are made with finite computational precision.

Introduction

After experiencing reward or punishment, agents update their estimates of the reinforcement value of relevant states and actions[1]. Over past decades, psychologists and neuroscientists have identified the update rules which describe value-guided learning in humans and other animals[2,3]. Meanwhile, statisticians and computer scientists have defined their normative properties[4]. In the lab, value-guided 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[5]. Value learning on this task can be modelled with a simple principle known as a delta rule[6]:

$$V_{t+1}^{i} = V_t^i + \left(R_t - V_t^i\right) \cdot \alpha \qquad [1]$$

Where V_t^i is the estimated value of bandit i on trial t, R_t is the payout obtained on trial t, and α is a learning rate in unity range. If α is sufficiently small, V^i is guaranteed to converge over time to the vicinity of expected value of bandit i.

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^i have been observed to fit human data from a 2-armed bandit task better, even after penalising for additional complexity[7-9]. 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[10]. 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[11-13]. 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.

Confirmation bias is a ubiquitous feature of human perceptual, cognitive and social processes and a longstanding topic of study in psychology[14]. 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[15]. 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[16]. However, this account is unsuited to explaining the differential updates to chosen and unchosen options in a bandit task, because in this case feedback for both options is freely displayed to the participant, and there is no overt data selection problem.

Here, we report simulations which reveal that paradoxically, confirmatory biases are beneficial in multi-armed tasks in the sense that under standard assumptions, they maximise the average

total reward (or rate of reward) for the agent. We find that this benefit holds over a wide range of settings, including both stationary and nonstationary bandits, across different epoch lengths, and under different levels of choice variability. These findings may explain why the humans tend to revise beliefs to a smaller extent when outcomes do not match with their expectations.

Results

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. Unless otherwise specified, in the following simulations we consider a multi-armed bandit task in which each bandit i is associated with a payout probability p^i sampled at intervals of 0.1 in the range [0.05; 0.95], that is a total of $\binom{10}{k}$ possible combinations of probabilities for a total of k bandits ($\binom{10}{2}$ = 45 for 2 bandits) (**Fig. 1a**). We consider an agent who chooses among bandits for 2^n trials (where n varies from 2-10; **Fig. 1b**) and, having received full feedback, updates the corresponding value estimate V^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)[13]. 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 $\alpha^C = \alpha^D$.

The agent makes decisions either following a hardmax rule (i.e. noiseless choices) or softmax rule (i.e. assuming Gaussian noise, with the level of noise determined by a temperature parameter β). We also consider cases where the reward probabilities are nonstationary; for example, in the 2-armed bandit case where the payout probabilities may reverse at regular intervals, or where they are sampled according to a random walk process (**Fig. 1c**). We conduct all simulations numerically, sampling the payout probabilities and experiment length(s) exhaustively, varying $\alpha^C$and α^D exhaustively, and noting the average reward obtained by the agent in each setting.

Our main finding is illustrated in **Fig. 2**. Here, we plot 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 the converse is true. 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 "neutral" setting for α. The area enclosed by black line in **Fig. 2a-b** indicate combinations of learning rates that yield rewards higher than the neutral 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.

Under the assumption that payout probabilities are stationary and decisions are noiseless (i.e. under a hardmax choice rule), the optimal α on trial t is $1/t$, such that the learning rate is

gradually annealed over time (because then V_t^i is equal to the average of rewards observed for option i until time t). We confirmed this by plotting the average reward under various temperature values for three models: one in which a single alpha 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**), a reversal of probability between the two bandits midway through the sequence (**Fig. 3b**), and for three such reversals (**Fig. 3c**), and for a random walk in which probabilities drift upwards or downwards on each trial (**Fig. 3d**). 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. Subsequently, we tested how the sequence length affected the relative advantage conferred by a confirmatory bias. In **Fig. 3e**, 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 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. 3f**).

Next, we returned to a previously published dataset that provided evidence for the confirmation bias in human decisions and asked whether participants' estimated levels of bias were those that maximised their return, conditional on their estimated decision noise. The results are shown in **Fig. 4**. The confirmation bias is expressed here in a single number, to which we refer as the normalized learning rate: $\alpha^{norm} = (\alpha^c - \alpha^D)/(\alpha^c + \alpha^D)$. The colour in the plot shows the average reward as a function of the confirmation bias (y-axis) and the level of noise (x-axis) in a simulation paralleling the experimental task[13]. Presented in this form, one can see that as decisions become more noisy, the range of confirmation bias giving high reward starts to include larger biases. For example, for a low level of noise $\beta = 0.1$, reward is on average highest for lower range of normalized learning rates $\alpha^{norm} \in (0.1, 0.5)$, while for a higher noise level $\beta = 0.7$, reward is maximized for a wider range of normalized learning rates $\alpha^{norm} \in (0.1, 0.9)$. The fits of individual participants from the experimental study are shown as white stars. As can be seen, estimated bias also tends to occupy a higher range with larger estimated noise. One interpretation of this finding is that humans adaptively scale their level of bias up or down to compensate for the stochasticity that corrupts their choices, as previously described[17,18].

These analyses 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 selected three parametrisations of the update rules and examined their consequences in more detail. The selected pairs of values for α^C and α^D are illustrated in **Fig. 5a** (symbols Δ, $\times$ and o). 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 ($\alpha^C = 0.45, \alpha^D = 0.05$). We chose a bandit setting in which $p^+ = 0.65$ and $p^- = 0.35$, although our results held over a wide range of equivalent choices.

For each update rule, we plotted the evolution of the value estimate for the more valuable bandit V^+ over trials (**Fig. 5b**) as well as aggregate choice accuracy (**Fig. 5c**). 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. 5c**). This is of course simply a restatement of the finding that biased policies maximise reward (see shading in **Fig. 5a**). However, perhaps more informative are the value estimates for V^+ under each update rule (**Fig. 5b**). 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 better 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, on each trial the agent obtains estimates V^1 and V^2, the softmax choice rule is equivalent to the following logistic function:

$$p^1 = \frac{1}{1+e^{\left(V^2-V^1\right)/\beta}} \qquad [2]$$

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 $V^1 - V^2$ away from zero in either the positive or the negative direction, thereby ensuring choice probabilities that were 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 worse option rather than the better, 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 lower option will be overvalued rather than the higher (see the bimodal distribution of value estimates in **Fig. 5b**, inset). Our simulations show that when this happens, the average reward is low, resulting in

bimodal distribution of rewards across simulations (inset in **Fig. 5a**). 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.

The analysis shown in **Fig. 5** illustrates why the benefit 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. 5c**). 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).

In our final analysis, 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 (DDM). This allows us to model not only the choice probabilities but also reaction times. To this end, we simulated a *reinforcement learning drift diffusion model* in which the drift rate was proportional to the difference in value estimates between the two bandits[19], which in turn depends on the update policy (confirmatory, disconfirmatory, or neutral). We employed the setting with 128 trials, used stable contingencies ($p^- = 0.35$ and $p^+ = 0.65$) and parametrized the drift diffusion process with ($v_{mod} = bound = \sqrt{0.1}$ and $noise = 0.1$, see ***Methods***) such that the signal-to-noise ratio is equivalent to that under the softmax temperature used in **Fig. 5** ($\beta = 0.1$).

When we plotted the overall accuracy of the model, the results closely resemble those from previous analyses, as is to be expected (**Fig. 6a**). When we examined simulated reaction times, we observed that confirmatory learning leads to faster decisions. This follows naturally from the heightened difference in values estimates for each bandit, as shown in **Fig. 5**. 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. 5c**). 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.

Discussion

Humans have been observed to exhibit confirmatory biases when choosing between stimuli or actions that payout with uncertain probability[11-13]. 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 almost all 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. 5**). This can have both positive and negative effects. The negative effect is that a 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 depend on the level of decision noise: within reasonable noise ranges the latter effect outweighs the former and performance benefits overall.

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[20]. 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[17]. 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[21]. 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[22]. 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[23]. Our findings are also consistent with previous reports that a bias towards optimism may be beneficial[9].

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. Implicitly, thus, we are committing to the idea that value-guided decisions may be irreducibly noisy even where exploration is not required[24]. 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[25]. Others have instead assumed a different form to the noise distribution, and modelled choice stochasticity with an "epsilon-greedy" policy (which introduces lapses to the choice process with probability epsilon). We did not include a simulation using an epsilon-greedy decision rule, but we think it very likely that it would be behave similarly to the hardmax rule.

Finally, we want to address the limitations of the present study. 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[26-28]. An analysis of these 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 (i.e. *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. Finally, the present study does not investigate the impact on the performance of other kinds of internal noise such like an update noise[25]. The latter, instead of perturbating the policy itself, perturbates 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. However, understanding how both noise sources may interact in simulations of the *confirmation* model remains of prime interest.

Methods

Simulation parameters

We simulated multi-armed bandit tasks with variable numbers of trials, as described in the main text. Ground truth bandit payout probabilities were either stable, reversing or drifting over time. Simulated agents received both factual and counterfactual feedback about their choices (i.e. they view the payout of both the chosen and unchosen options) on each trial, with outcomes being either 0 or 1. In the two-armed case, the initial probabilities p^1 and p^2 of obtaining a reward from each arm are defined in steps of 0.1 within the interval $[0.05, 0.95]$ in three out of four volatility conditions (*stable*, *1 reversal*, *3 reversals*). In the *stable* condition, these probabilities stay the same across the whole learning period whereas in the *1 reversal* condition, they are reversed to $1 - p$ at the midpoint of the period and in the *3 reversals*

condition, they are switched to $1 - p$ three times, after 0.25, 0.5 and 0.75 of the trials have elapsed.

In the fourth *random walk* condition, probabilities are randomly initialized probabilities and then drift over trials as follows:

$$p_{t+1} = p_t + \kappa \cdot (0.5 - p_t) + \mathcal{N}(0, \sigma^2)$$

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 probabilities are sampled (here set to $\sigma = 0.02$). For all but the *random walk* condition, we tested all the possible combinations of initial probabilities with p^1 and p^2 defined between 0.05 and 0.95 by increment of 0.1 and $p^1 \neq p^2$ (that is 45 probability pairs in the case of n = 2), and unless otherwise noted, results are averaged across these cases.

The simulations also vary in terms of period length with the number of trials calculated as 2^n with n defined as integers in the interval $[2, \ldots, 10]$. All simulations are 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.

Reinforcement learning models

All the simulations of the main analysis are performed using what is called the *confirmation* model in Palminteri et al.[9]. The model adapts the standard delta-rule update but involves two different learning rates, one for confirmatory prediction errors (i.e. positive for the chosen option and negative for the unchosen option) and one for disconfirmatory predictions errors (i.e. negative for the chosen option and positive for the unchosen option). Then at each time t, if the agent chooses the option 1, the model updates the values V^1 and V^2 of the chosen and unchosen options respectively, such that:

$$V_{t+1}^1 = V_t^1 + \begin{cases} \alpha^C \cdot PE_t^1, & PE_t^1 > 0 \\ \alpha^D \cdot PE_t^1, & PE_t^1 < 0 \end{cases}$$

and

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

with PE^i being the prediction error for bandit i on trial t for the chosen option and calculated as:

$$PE_t^i = R_t^i - V_t^i$$

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 0.1, that is 19^2 learning rate combinations. Note that for $\alpha^C = \alpha^D = \alpha$, the model amounts to a standard delta-rule model with a unique

learning rate α, used for all types of predictions errors (positive and negative) and values (chosen and unchosen), such that V^i is updated as follow:

$$V_{t+1}^i = V_t^i + \alpha \cdot PE_t^i$$

with PE_t^i defined as above.

Then the simulations performed with these particular learning rates combinations serve as a benchmark to compare *biased* and *unbiased* learning.

Decaying learning rate model. In addition to the *confirmation* model with all learning rates combinations, a decaying learning rate model is used as a second benchmark in the stable condition when we relax the assumption of fixed learning rate. The update in this model is similarly defined as in the aforementioned *unbiased* model except that the learning rate α is defined such that:

$$\alpha = \frac{1}{t}$$

with t being the trial number such that α decays over trials.

Decision Policies

All models presented above make decisions through either a *hardmax* or a *softmax* policy. The former is a noiseless policy selecting deterministically the arm associated to the highest value whereas the latter is a probabilistic action selection process associating to each of the arm of probability p of being selected based on their respective values such that:

$$p_t^a = \frac{e^{V_t^a/\beta}}{\sum_{i=1}^{n} e^{V_t^i/\beta}}$$

with n, the number of arms and β the temperature of the *softmax* function, the higher the temperature, the more random the decision is.

The simulations were performed with different β defined on the interval [1/10, 10] such that:

$$\beta \in [0.1, \ldots, 1] \cup [1/0.1, \ldots, 1/1]]$$

Drift diffusion model

The last part of the main analysis is performed adding a drift diffusion process to the *confirmation* model, in order to estimate decision reaction times at each trial t from the difference in values at that moment and make decision as a result.

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:

$$x_{i+1} = x_i + v_t * dt \ + \ \varepsilon * \sqrt{dt} * \mathcal{N}(0,1)$$

with x_0, the initial evidence defined as:

$$x_0 = \frac{a}{2}$$

and dt set to 0.001 and ε to 0.1. The drift rate v_t, is linearly defined from the difference in values such that:

$$v_t = v_{mod} * (V_t^+ - V_t^-)$$

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 [29] is given by:

$$p^+ = \frac{1}{1 + e^{-av_t/\varepsilon^2}}$$

The above probability is equal to that in Equation 2 if $av_{mod}/\varepsilon^2 = 1/\beta$. Thus, we set $v_{mod} = a \ = \sqrt{0.1}$.

The values are updated exactly the same way as in the *confirmation* model, described above.

Summary of parameters

Description	Symbol	Range of value tested
Confirmation Model		
Confirmatory learning rate	α^C	$[0.05, 0.10, \ldots, 0.95]$
Disconfirmatory learning rate	α^D	$[0.05, 0.10, \ldots, 0.95]$
Softmax Temperature	β	$[0.1, \ldots, 1] \cup [1/0.1, \ldots, 1/1]$
Policy	n.a.	$[hardmax, softmax]$
RLDDM		
Drift rate scaling parameter	v_{mod}	$\sqrt{0.1}$
Threshold	a	$\sqrt{0.1}$
Noise	ε	0.1
Time unit	dt	0.001
Bandit Task		
First option probability	p_1	$[0.05, 0.10, \ldots, 0.95]$
Second option probability	p_2	$[0.05, 0.10, \ldots, 0.95]$
Trial number	t	$[4, 8, \ldots, 1024]$
Volatility type	n.a.	$[stable, 1\ reversal, 3\ reversals, random\ walk]$

Table 1. Simulations parameters. The table summarizes the different algorithm's and task's parameters and their range of definition.

References

1 Thorndike, E. L. Some Experiments on Animal Intelligence. *Science* **8** (1898).

2 Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiol Rev* **95**, 853-951, doi:10.1152/physrev.00023.2014 (2015).

3 Dayan, P. & Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* **8**, 429-453 (2008).

4 Sutton, R. & Barto, A. *Reinforcement Learning*. (MIT press, 1998).

5 Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876-879 (2006).

6 Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds Black A.H. & W.F. Prokasy) 64-99 (Appleton Century Crofts, 1972).

7 Gershman, S. J. Do learning rates adapt to the distribution of rewards? *Psychon Bull Rev* **22**, 1320-1327, doi:10.3758/s13423-014-0790-3 (2015).

8 Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J Neurosci* **32**, 551-562, doi:10.1523/JNEUROSCI.5498-10.2012 (2012).

9 Caze, R. D. & van der Meer, M. A. Adaptive properties of differential learning rates for positive and negative outcomes. *Biol Cybern* **107**, 711-719, doi:10.1007/s00422-013-0571-5 (2013).

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

11 Chambon, V. *et al.* Choosing and learning: outcome valence differentially affects learning from free versus forced choices. *BiorXiv preprint* (2019).

12 Schuller, T. *et al.* Decreased transfer of value to action in Tourette syndrome. *Cortex* **126**, 39-48, doi:10.1016/j.cortex.2019.12.027 (2020).

13 Palminteri, S., Lefebvre, G., Kilford, E. J. & Blakemore, S. J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Comput Biol* **13**, e1005684, doi:10.1371/journal.pcbi.1005684 (2017).

14 Nickerson, R. S. Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology* **2**, 175-220 (1998).

15 Groopman, J. *How Doctors Think*. (Mariner Books, 2007).

16 Oaksford, M. & Chater, N. Optimal data selection: revision, review, and reevaluation. *Psychon Bull Rev* **10**, 289-318, doi:10.3758/bf03196492 (2003).

17 Li, V., Herce Castanon, S., Solomon, J. A., Vandormael, H. & Summerfield, C. Robust averaging protects decisions from noise in neural computations. *PLoS Comput Biol* **13**, e1005723, doi:10.1371/journal.pcbi.1005723 (2017).

18 Spitzer, B., Waschke, L. & Summerfield, C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat Hum Behav* **1**, doi:10.1038/s41562-017-0145 (2017).

19 Pedersen, M. L., Frank, M. J. & Biele, G. The drift diffusion model as the choice rule in reinforcement learning. *Psychon Bull Rev* **24**, 1234-1251, doi:10.3758/s13423-016-1199-y (2017).

20 Summerfield, C. & Tsetsos, K. Do humans make good decisions? *Trends Cogn Sci* **19**, 27-34, doi:10.1016/j.tics.2014.11.005 (2015).

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

22 Tsetsos, K. *et al.* Economic irrationality is optimal during noisy decision making. *Proc Natl Acad Sci U S A* **113**, 3102-3107, doi:10.1073/pnas.1519157113 (2016).

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

24 Renart, A. & Machens, C. K. Variability in neural activity and behavior. *Curr Opin Neurobiol* **25**, 211-220, doi:10.1016/j.conb.2014.02.013 (2014).

25 Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S. & Wyart, V. Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nat Neurosci* **22**, 2066-2077, doi:10.1038/s41593-019-0518-9 (2019).

26 Worthy, D. A., Pang, B. & Byrne, K. A. Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Front Psychol* **4**, 640, doi:10.3389/fpsyg.2013.00640 (2013).

27 Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol Rev* **126**, 292-311, doi:10.1037/rev0000120 (2019).

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

29 Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* **113**, 700-765 (2006).

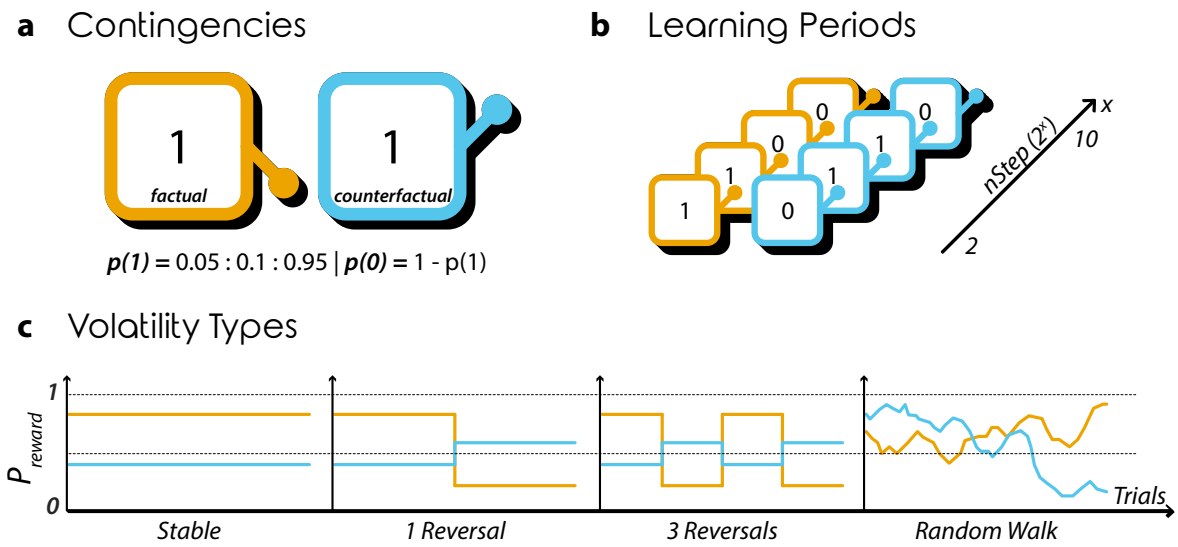

Figure 1. Simulation Setup. a. Reward contingencies. The illustration represents the chosen (orange) and unchosen (blue) bandits each with a feedback signal (central number). Below, we state the range of possible outcomes and probabilities. **b**. Learning Periods. The illustration represents the different 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.

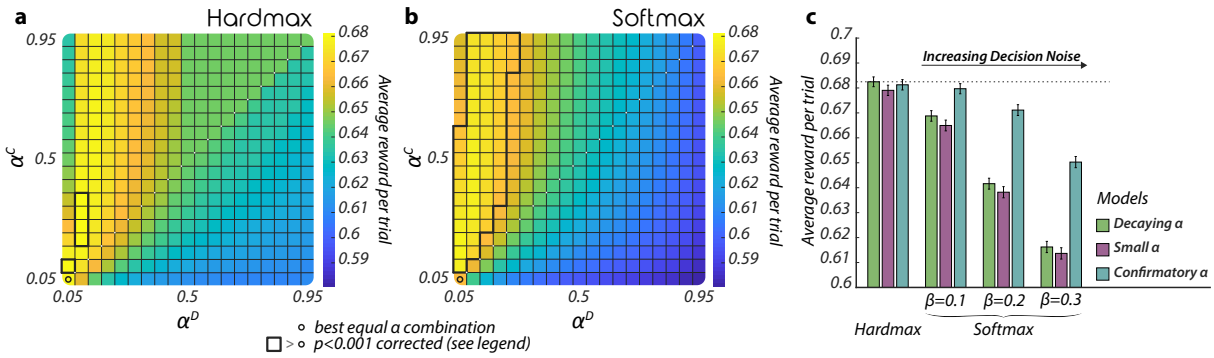

Figure 2. Dependence of reward on learning rate and decision noise in a stable environment. **a** and **b**. Average reward for all learning rate combinations. The heatmaps represent the per trial average reward for combinations of α^C (y-axis) and α^D (x-axis), averaged across all reward contingencies 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. 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.

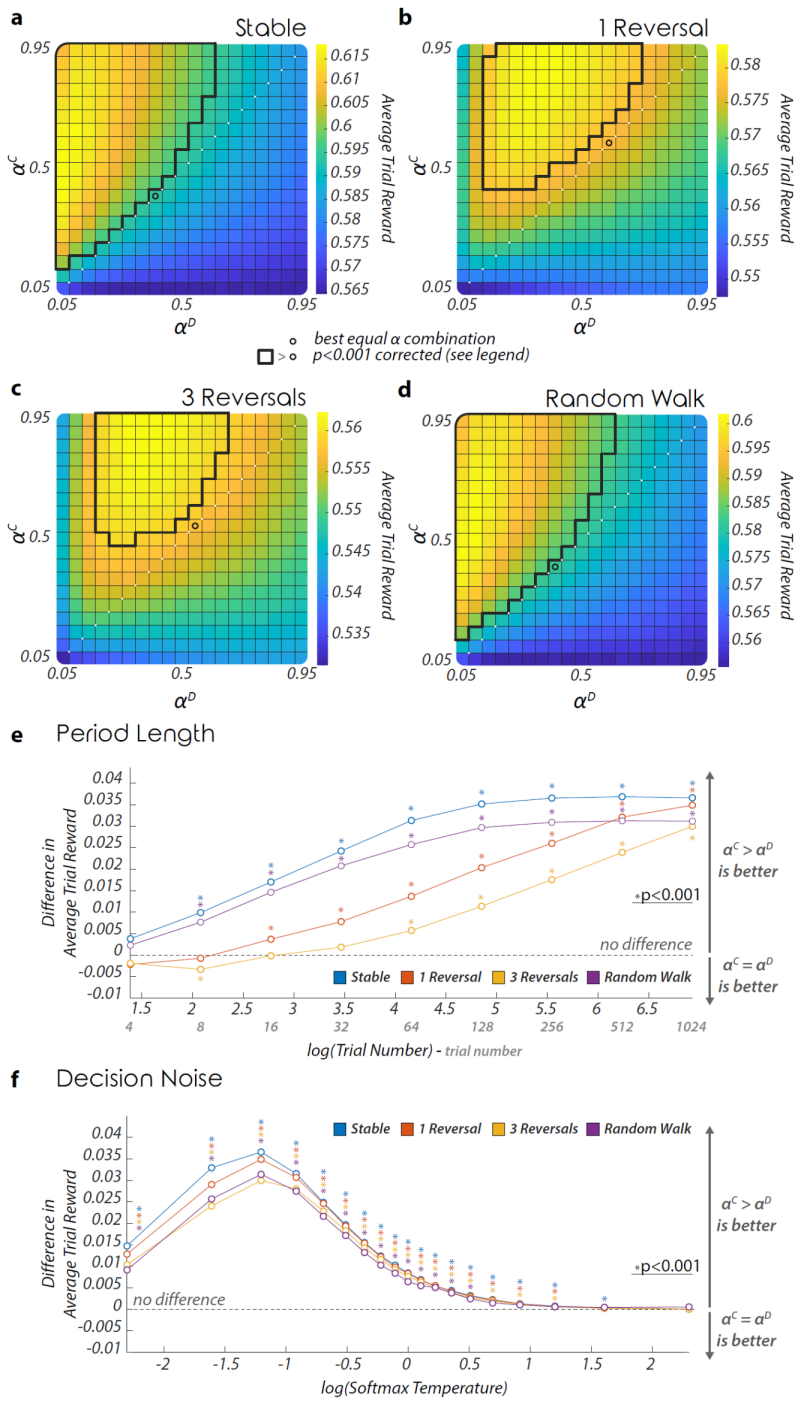

Figure 3. Dependence of reward on learning rate and decision noise in a different environments. a, b, c and **d**. The heatmaps represent the per trial average reward for combinations of α^C (y-axis) and α^D (x-axis) given a *softmax* policy ($\beta = 0.3$). The performance is averaged across all reward contingencies, period lengths and 1000 agents in the stable condition (**a**), 1 reversal condition (**b**), 3 reversals condition (**c**) or 100000 agents in the random walk condition (**d**). 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. **e**. Effect of period length on reward. The line plot represents the difference in average reward between the confirmation model (with the

per period best confirmatory learning rate combination) 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. **f**. 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 rates combination) 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.

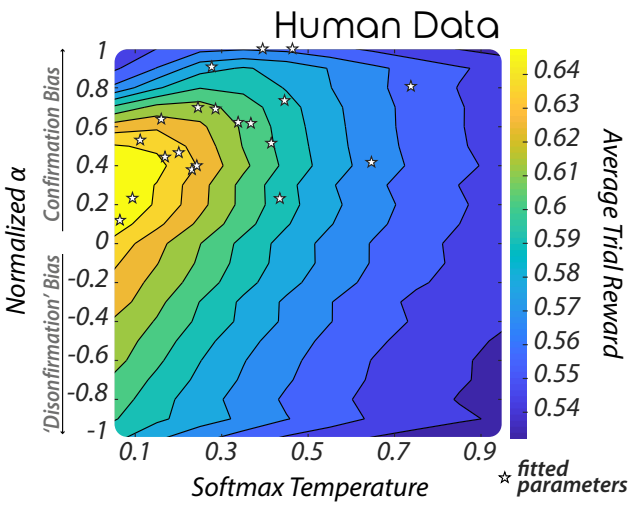

Figure 4. Parameters estimated from Human Data. The contour plot represents the per trial average reward simulated with the confirmation model as a function of the normalized difference between learning rates on the y-axis (a positive value indicates a confirmatory combination of learning rates whereas a negative value indicates the converse) and the *softmax* temperature β on the x-axis. The normalized learning rates was computed as $\alpha^{norm} = (\alpha^c - \alpha^D)/(\alpha^c + \alpha^D)$ for all combinations of learning rates α^C and α^D defined in the range $[0.05, 0.95]$ with increments of 0.1 and binned in 21 categories defined from by increment of 0.1 ($-0.05 < \alpha^{norm} < 0.05$, being the middle category). Stars represent the pairs of parameters (α^{norm} and β) fitted on real subjects (the fitting procedure is detailed in the original study). The conditions simulated follow the protocol of the experiment from which the human data come from (Conditions 1 and 2: $p^+ = 0.75$ and $p^- = 0.25$, condition 3: $p^+ = p^- = 0.50$, and condition 4: $p^+ = 0.87$ and $p^- = 0.13$ with a reversal of probabilities in the middle. All conditions compound 24 trials and were repeated twice). Results are obtained by simulating 1000 agents on this task.

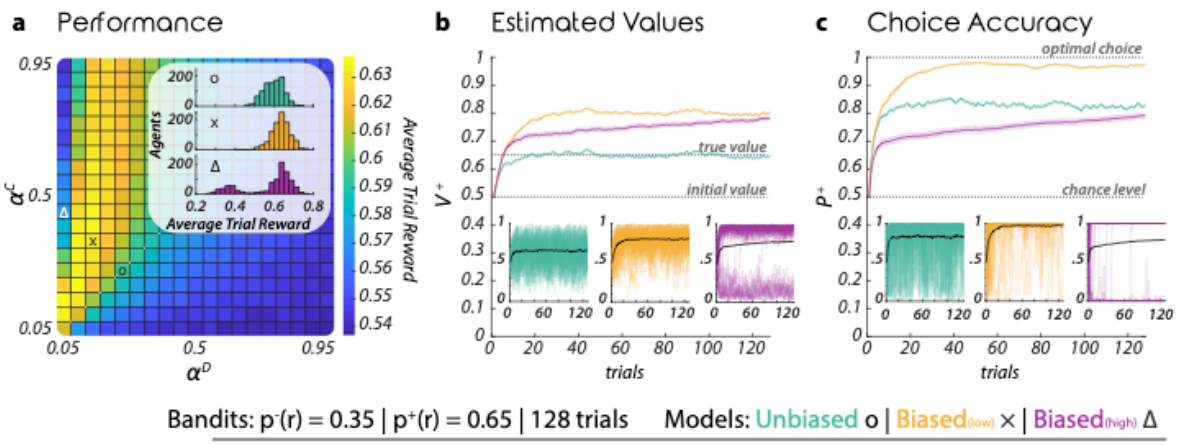

Figure 5. Mechanism by which confirmation bias tends to increase reward. a. Average reward and reward distributions for different levels of confirmation bias. The heatmap represents the per trial average reward of the confirmation model for all learning rates combinations (confirmatory learning rates 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 ($p^{-}(r) = 0.35$ and $p^{+}(r) = 0.65$) and are averaged across agents. The three signs inside the heatmap (Δ, x 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" ($\alpha^{C} = \alpha^{D} = 0.25$), "biased (low)" ($\alpha^{C} = 0.35$ and $\alpha^{d} = 0.15$) and "biased (high)" ($\alpha^{C} = 0.45$ and $\alpha^{D} = 0.05$). 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" ($\alpha^{C} = \alpha^{D} = 0.25$), "biased (low)" ($\alpha^{C} = 0.35$ and $\alpha^{D} = 0.15$) and "biased (high)" ($\alpha^{C} = 0.45$ and $\alpha^{CD} = 0.05$). 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.

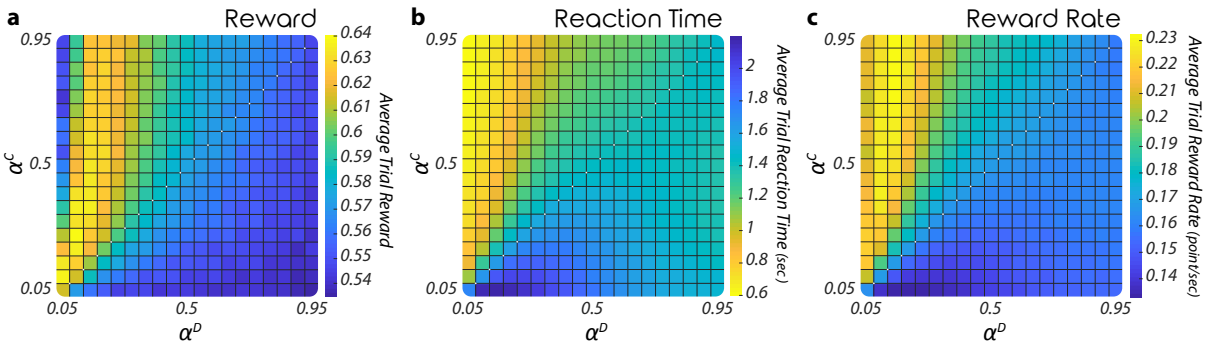

Figure 6. Effect of confirmation bias on reward rate. a. The heatmap represents the per trial average reward simulated with the confirmation RLDDM for all learning rates combinations (confirmatory learning rates are represented on the y-axis whereas disconfirmation learning rates are represented on the x-axis). The rewards concern the stable condition with 128 trials and asymmetric contingencies ($p^{-}(r) = 0.35$ and $p^{+}(r) = 0.65$) 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.