Learning asymmetry or perseveration? A critical re-evaluation and solution to a pervasive confound

Juan Vidal-Perez[1,2]*, Raymond J. Dolan[1,2], Rani Moran[1,3]*

Affiliations:

[1] Max Planck Centre for Computational Psychiatry and Ageing, University College London, Russell Square House, WC1B 5EH, England

[2] Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, England

[3] Department of Psychology, School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

***Corresponding authors**. Email: rani.moran@gmail.com, juan.perez.21@ucl.ac.uk

ABSTRACT

A central challenge in cognitive science is to distinguish between multiple processes that can result in similar behaviours. In reinforcement learning (RL), one prominent example concerns two potential drivers of choice repetition: (i) a confirmation-bias learning asymmetry, in which agents learn more from outcomes that confirm their choices, and (ii) choice perseveration, an outcome-independent tendency to repeat past choices. Evidence for asymmetric learning has typically relied on computational models that control for perseveration, or specific behavioural markers designed to reveal asymmetric learning. Here, we show both these approaches have critical flaws and can spuriously detect learning asymmetries even in perseverative, symmetric-learning agents. To address this, we introduce a novel statistical test that distinguishes genuine learning asymmetries from spurious effects. Applying this test to a large dataset spanning ten published experiments, we find that some previously reported confirmation biases are fragile, albeit others remain robust even at a meta-analytic level. Finally, we propose a new task design that can yield a more valid qualitative signature of confirmation bias. We suggest our approach provides a reliable framework for disentangling processes underlying choice repetition, while providing tools for the wider research community that can minimize potential spurious effects arising from process mimicry and biased parameter estimation.

INTRODUCTION

Reinforcement learning (RL) models are an indispensable tool for dissecting mechanisms of human decision-making. RL models formalize learning as a trial-and-error process wherein an agent updates the expected value of different options based on a difference between expected and received rewards, a quantity known as prediction error (1). While standard models assume this updating process is symmetric (i.e., occurring to an equal extent for positive and negative prediction errors), an emerging insight has been that human learning is often asymmetric.

Studies, primarily in the context of bandit tasks, show that better-than-expected outcomes (confirming a choice) drive learning more readily than worse-than-expected outcomes (disconfirming a choice) (2–6). This mirrors other learning asymmetries in learning, including greater belief updates following positive feedback (optimism bias) (7,8) or feedback that aligns with our identity or prior beliefs (confirmation bias) (9–12). Such learning biases are suggested to provide a mechanistic account for a wide range of phenomena, ranging from disorders like depression (13–15) through to polarized belief formation (16,17) and cognitive processes such as transitive inference (18). Recent work has suggested learning asymmetries are operative at the single-neuron level and form the basis for distributional RL theories (19–21).

To infer learning asymmetries in behavior, choice-repetition patterns are frequently invoked. Models that assume symmetric learning frequently predict choice-repetition rates that are either higher or lower than empirically observed. By contrast, asymmetric outcome-based learning models naturally account for such discrepancies. Overweighting positive feedback for a chosen option, renders an agent more likely to repeat that choice, creating a self-reinforcing loop (22), while overweighting negative feedback predicts increased choice alternation. However, gradual choice-perseveration (hereafter referred to simply as “perseveration”), a tendency to repeat previous choices regardless of their outcomes, provides an alternative account for choice hysteresis/inertia. Here a positive perseverative tendency can arise out of habit formation (23). Conversely, novelty seeking, a drive to explore less-frequently chosen options, leads to choice alternation or “negative perseveration” (24,25). Consequently, to better understand the mechanisms generating choice behavior (e.g., whether these are choice or outcome-based) it is critical to dissociate learning asymmetries from perseveration. Here, we focus on how we can dissociate confirmation bias from gradual choice-perseveration, a topic that has attracted interest and debate in recent years.

Researchers studying learning asymmetries face the problem of *model mimicry*, whereby perseverative behavior can masquerade as a learning asymmetry (22). In other words, when an agent learns symmetrically but is also subject to a perseverative influence then fitting their behavior with a model that ignores the latter tendency will lead to a systematic error. The model will incorrectly attribute outcome-independent choice repetition to an outcome-dependent learning bias, resulting in a spurious, by-construction, estimate of a learning asymmetry. From this perspective, perseveration is a nuisance process that must be controlled for in order to uncover a genuine contribution from a learning asymmetry (i.e., the primary focus of interest).

Researchers typically followed two main strategies to deal with this problem. Firstly, they use a hybrid model, that includes simultaneous contributions from learning asymmetries and perseveration. Here the goal is to isolate a learning asymmetry that "survives" a statistical control for perseveration. Secondly, researchers have developed behavioral markers or ‘signatures’, which assume to selectively indicate learning asymmetries while being insensitive to perseveration. Here, we reevaluate these strategies in turn, highlighting pitfalls with current approaches and proposing solutions.

From the perspective of modeling, a critical assumption is that hybrid model yields unbiased estimates of learning asymmetries by fully controlling for perseveration. Notably, previous reports of a significant confirmation bias (measured ignoring perseveration) become non-significant when reanalyzed using hybrid models (26). However, here, we show this critical unbiasedness assumption is not necessarily correct and that the prior application of the hybrid model yields biased learning-asymmetry estimates, risking spurious conclusions. Indeed, perseverative agents, with symmetric value updating, can systematically be misidentified as having a learning asymmetry (spurious by construction). We show this can emerge particularly from use of designs with limited number of trials per-participant (typically the case in many experimental designs) and is further amplified when the fitting procedure uses perseveration-shrinking parameter priors. While this can be mitigated when parameter-priors are removed, it disappears only with an unfeasibly large number of trials. Additionally, we also reexamine behavioral signatures previously proposed as indicators of learning asymmetries, highlighting critical limitations. We show these signatures are sensitive to perseveration in ways that either invalidate them or substantially limit their practical utility.

The difficulty of reliably dissociating learning asymmetries from perseveration inspired us to develop a novel, empirically-grounded, computational solution. We introduce a bespoke bootstrapping-based statistical test that controls for biased estimates within the hybrid model. This provides a rigorous method to test for genuine learning asymmetries by allowing for the rejection of a null hypothesis, where behavior is driven *only* by perseveration and symmetric learning. Applying this method to four prominent studies drawn from the literature (spanning 10 experiments), we show that evidence for learning asymmetries (specifically positivity/confirmation biased) is more nuanced than previously thought. Indeed, several studies, which in previous hybrid-model-based analyses were considered to provide evidence for a confirmation bias, fail to do so under our proposed novel hypothesis test. Reassuringly, we still find evidence for a significant positive learning asymmetry within individual studies and at the meta-analytic, across-study, level. Finally, we present a novel task design capable of providing a qualitative behavioral signature that directly implicates learning-asymmetry thus providing a more robust framework for studying learning biases, minimizing a potential subtle but significant methodological confound. Finally, while we focus on dissociating learning asymmetries from perseveration in reinforcement learning, the issues we highlight have much broader relevance. In particular, our findings highlight more general challenges such as process-mimicry and biased parameter estimation that may contribute to spurious conclusions.

RESULTS

A Hybrid Model Reduces, but Does Not Eliminate, Measured Confirmation Bias in Empirical Data

A central challenge in modeling reinforcement learning is to dissociate the influence of asymmetric value updating (i.e., confirmation bias) from that of choice perseveration (22,26–28). It has been argued that hybrid models, which incorporate parameters for both mechanisms, can effectively disentangle their respective contributions to behavior (26). Here, we scrutinize this claim by focusing on a prominent hybrid model architecture proposed by Palminteri et al. (27).

In this model, the values (Q-values) of available options are updated based on prediction errors (δ_t), but the learning rate applied depends on whether an outcome confirms or disconfirms the choice. For a chosen option C, the Q-value is updated as follows:

$$Q_{t+1}(C) = \begin{cases} Q_t(C) + \ \alpha_c \cdot \delta_t(C)\ if\ \delta_t(C) > 0 \\ Q_t(C) + \ \alpha_d \cdot \delta_t(C)\ if\ \delta_t(C) < 0 \end{cases}$$

where $\delta_t(C) = R_t(C) - Q_t(C)$ is the prediction error. Better-than-expected outcomes ($\delta_t(C) > 0$) are considered confirmatory and are learned from with rate α_c, while worse-than-expected outcomes ($\delta_t(C) < 0$) are disconfirmatory and learned from with rate α_d.

Research on learning asymmetries typically uses bandit tasks with either partial feedback (outcome for the chosen option only) or full feedback (outcomes for both chosen and unchosen options). This distinction is critical, as full feedback allows for a more complete test of confirmation bias, which predicts a specific, reversed learning pattern for unchosen options: a better-than-expected outcome for this option is treated as disconfirmatory, as it suggests the agent made the wrong choice. Accordingly, the unchosen option value is updated as follows:

$$Q_{t+1}(U) = \begin{cases} Q_t(U) + \ \alpha_c \cdot \delta_t(U)\ if\ \delta_t(U) < 0 \\ Q_t(U) + \ \alpha_d \cdot \delta_t(U)\ if\ \delta_t(U) > 0 \end{cases}$$

Across both partial and full feedback designs, the degree of confirmation bias is quantified from the fitted learning rate parameters. We assess this using two standard metrics: the absolute confirmation bias, calculated as the simple difference ($\alpha_c - \alpha_d$), and the normalized confirmation bias, which accounts for the overall learning rate, calculated as $(\alpha_c - \alpha_d)/(\alpha_c + \alpha_d)$.

In parallel, the model tracks choice perseveration using a choice trace, $C_t(i)$, for each option i. This trace is updated on every trial:

$$C_{t+1}\ (i) = \begin{cases} C_t(i) + \tau \cdot \big(1 - C_t(i)\big)\ if\ i = C \\ C_t(i) + \tau \cdot \big(0 - C_t(i)\big)\ if\ i = U \end{cases}$$

The parameter τ governs the accumulation rate of choice traces, in practice acting as a learning rate for choice history.

Finally, both the learned values and the choice traces are injected into a softmax decision rule to compute the probability of choosing one option over another:

$$p(choose\ A\ over\ B) = \frac{1}{1 + e^{-\beta(Q_A - Q_B)} + e^{-\phi(C_A - C_B)}}$$

Here, β is the inverse temperature scaling the influence of Q-values, and ϕ is the inverse temperature parameter scaling the choice trace.

We first applied this hybrid model to four publicly available datasets (2–4,26), spanning 10 bandit-task experiments (see “Methods: Analysed datasets”), and compared it to a Pure Confirmation Bias (Pure CB) model that allows for asymmetric updating but not perseveration ($\phi, \tau = 0$) (Fig. 1 top and middle arms). Importantly, for comparability with previous results we followed the same fitting approach as used in these previous studies (see “Methods: Maximum a Posteriori estimation (MAP)”). Replicating the findings of Palminteri et al. (27), the Pure CB model detected a large confirmation bias ($\alpha_c - \alpha_d > 0$) (t-test, p<0.05 in all experiments). This bias was still evident in most experiments when fitting the hybrid model (p<0.05 in all experiments except for P1,C1,C3 and S1a), but in actuality was significantly reduced (paired t-test, p<0.05 in all experiments except for P2, C1, C3, and C4) due to choice perseveration parameters absorbing a portion of the behavioral variance (see Fig. 2a). The same conclusion is reached when examining the confirmation bias normalized by the overall degree of learning, $(\alpha_c - \alpha_d)/(\alpha_c + \alpha_d)$ (Pure CB, all p’s<0.05; Hybrid model, all p’s<0.05 except for C3 and S1a; significant decrease in confirmation bias, p<0.05 for all experiments except for C3; see. Fig.2b), another standard metric for value learning asymmetries (6,27).

Under the assumption that the hybrid model fully controls for perseverative influences, thereby allowing unbiased estimation of learning asymmetries, this supports a conclusion that a genuine confirmation bias survives this control in most datasets. However, we next show that this assumption is incorrect, and that perseveration with symmetric learning rates (PSL) can produce artifactual confirmation biases.

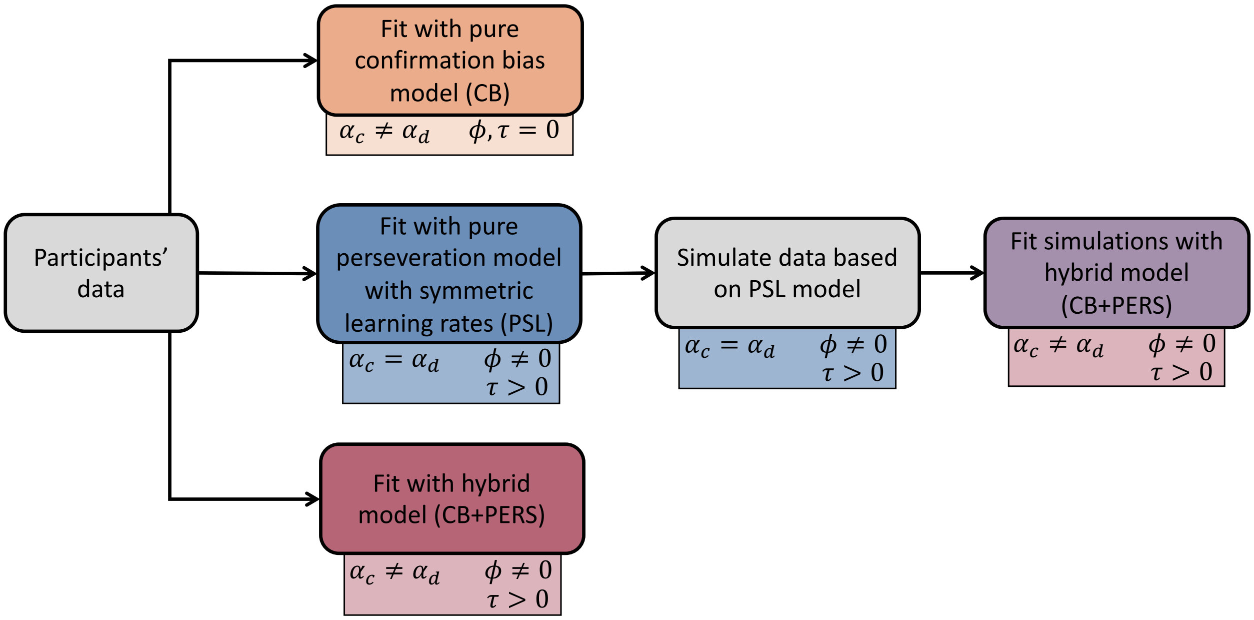

Fit with pure confirmation bias model (CB)
$\alpha_c \neq \alpha_d$ $\phi, \tau = 0$
Participants’ data
Fit with pure perseveration model with symmetric learning rates (PSL)
$\alpha_c = \alpha_d$ $\phi \neq 0$ $\tau > 0$
Simulate data based on PSL model
$\alpha_c = \alpha_d$ $\phi \neq 0$ $\tau > 0$
Fit simulations with hybrid model (CB+PERS)
$\alpha_c \neq \alpha_d$ $\phi \neq 0$ $\tau > 0$
Fit with hybrid model (CB+PERS)
$\alpha_c \neq \alpha_d$ $\phi \neq 0$ $\tau > 0$

Figure 1: Schematic of the model comparison and simulation procedure. *The figure illustrates the full analysis pipeline applied to data from each of 10 bandit-task experiments (top right box). We fitted participant data with three different models: (i) a Pure Confirmation Bias (CB) model, which allows for asymmetric value updating but not choice perseveration ($\phi = 0$) (**top arm**); (ii) a perseveration with symmetric learning rates (PSL) model, which enforces symmetric value updating ($\alpha_c = \alpha_c$) but allows for choice perseveration (**middle arm, blue box**); and (iii) a Hybrid model, allowing for both asymmetric updating and choice perseveration (**bottom arm**). To quantify artifactual confirmation bias due to choice perseveration, we implemented a simulation and re-fitting procedure where , we first obtained best-fitting parameters from the PSL model fits (**blue box in middle arm**). We then used these parameters to simulate new datasets (**gray box in middle arm**), creating a ground truth with perseveration but no true confirmation bias. Finally, these simulated datasets were re-fitted with the Hybrid model (**violet box in middle arm**) to measure the magnitude of any spurious confirmation bias that emerged.*

Choice Perseveration Masquerades as Spurious Confirmation Bias

Here, we took a "devil's advocate" approach to demonstrate that, using the above method, a perseverative agent (without any learning asymmetries) could be misidentified as having a confirmation bias. Previous work by Sugawara and Katahira (26) suggested that such a hybrid model can accurately dissociate the effects of asymmetric value updating (a difference between α_c and α_d) from those of choice perseveration (a positive ϕ). However, this conclusion was based on simulations of a particular task design and set size, where model parameters were drawn from specific prior distributions, leaving open a question regarding generalizability. We show that this conclusion does not generally hold across the diverse task designs and empirically-derived parameter distributions represented in our datasets.

In pursuing our goal, we used an alternative, empirically-grounded simulation and re-fitting procedure (Fig. 1, bottom arm). First, to estimate each participant's degree of choice perseveration under an assumption that their learning is truly symmetrical, we fit their data with a perseverative symmetrical learning (PSL) model, enforcing symmetric value updating ($\alpha_c = \alpha_d$) but allowing for choice perseveration ($\phi \neq 0, \tau > 0$). Second, we used the parameters estimated from each participant to simulate new behavioral datasets from PSL agents. These simulations represent a ground truth scenario where behavior is generated by symmetric updating and choice perseveration, with no underlying confirmation bias. Finally, we fit these simulated datasets with the full hybrid model (which allows $\alpha_c \neq \alpha_d$). If the hybrid model allows for an unbiased estimation of learning asymmetries, then no confirmation bias should be detected in these fits. In other words, any significant positive difference found between α_c and α_d in these simulations represents, by construction, a spurious confirmation bias (i.e., an artifact of model fitting where the behavioral patterns generated by perseveration are misattributed to asymmetric value updating). We applied this approach to the publicly available datasets detailed in the previous section. To minimize prediction-noise, we used a large number of simulations (1001 per participant), which provide high statistical power to detect small, but consistent, learning-asymmetries (see "Methods: Model simulations" for simulation specifications).

We found a significant artifactual confirmation bias in nearly all experiments for both the absolute difference in learning rates (all p's<0.05; except for C4, p=0.07) and in the normalized bias metric (all p's<0.05) (see Fig.2a-b, violet points). This demonstrates that contributions from choice perseveration can be systematically misattributed to asymmetric value updating (also see SI 1.1), highlighting a confound that persists even when explicitly accounting for both processes. These findings hold even when we use a more flexible model with separate learning rates for chosen and unchosen options (see SI 1.2).

We also examined how the presence of the asymmetric learning parameters in the hybrid model affected the estimation of the perseveration parameter, ϕ. Fitting the actual participant data, we show that the estimated perseveration weight (ϕ) was significantly greater for the PSL model than for the hybrid model (Fig. 2c, red vs. blue points). Complementing this, our simulation analysis showed that when fitting the PSL simulations with the hybrid model, the recovered ϕ parameter was systematically underestimated compared to its true generative value (Fig. 2c, violet vs. red points).

Overall, these results point to a critical tradeoff within the hybrid model: the emergence of a spurious confirmation bias comes at the cost of underestimating the true contribution of choice perseveration. The model incorrectly partitions the variance from a single perseverative mechanism into two separate, and consequently misestimated, parameters. However, even in the absence of perseveration we find a much smaller spurious confirmation bias (see SI 1.3). We also found this tradeoff does not operate in the opposite direction: when simulating an agent with a true confirmation bias but no perseveration, the hybrid model does not attribute some of the variance to a spurious perseveration parameter (see SI 1.4).

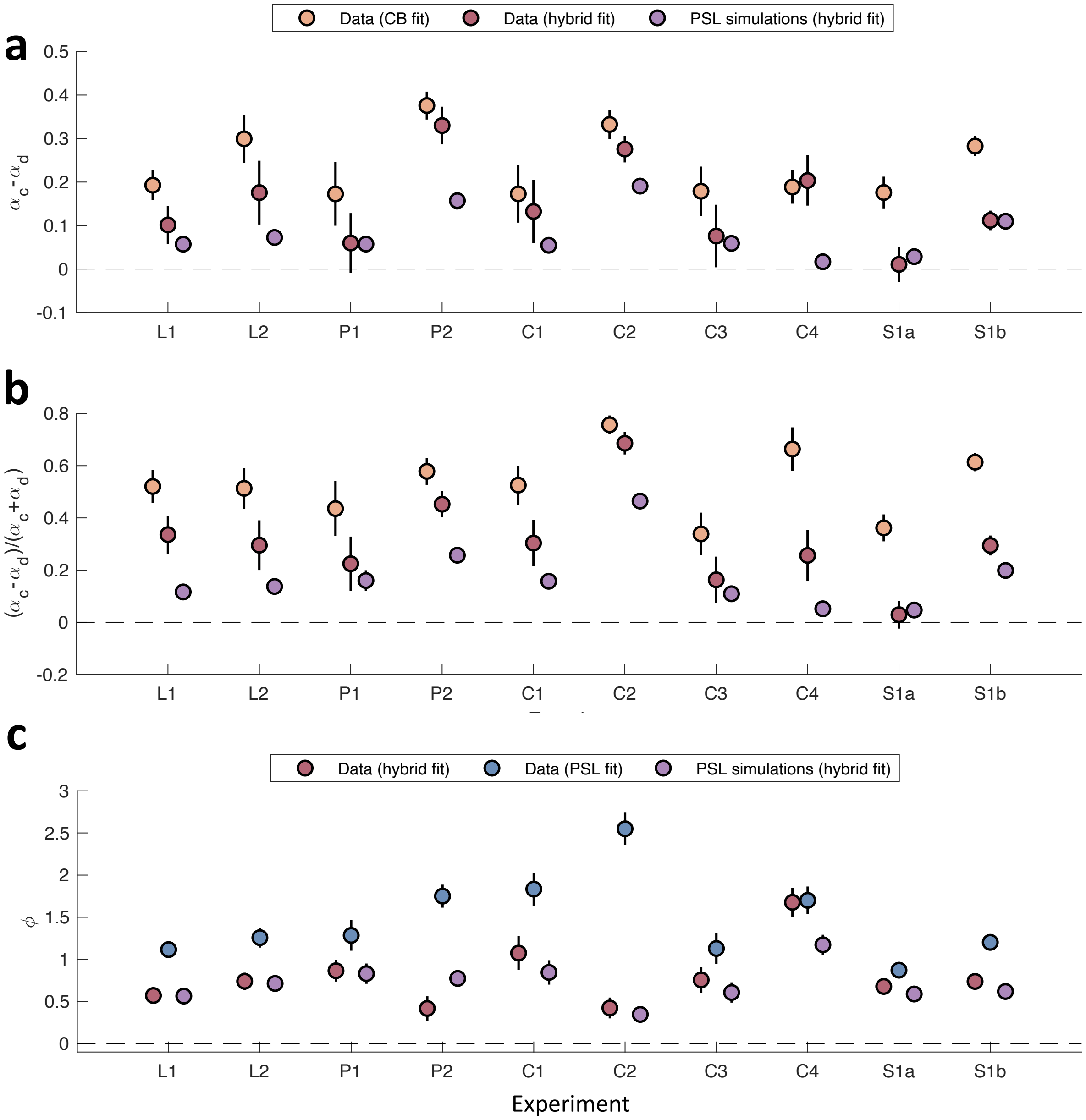


Figure 2: Confirmation bias and perseveration parameters estimated from empirical data and simulations. (a) *Absolute confirmation bias ($\alpha_c - \alpha_d$) estimated from fitting participant data with the Pure CB model (**yellow points**) and the Hybrid model (**red points**). Additionally, **violet points** show the spurious confirmation bias recovered when fitting PSL simulations with the Hybrid model. A spurious confirmation bias is detected across most experiments even when the generative model has no learning asymmetry. **(b)** Same as (a), but for the normalized confirmation bias, defined as $(\alpha_c - \alpha_d)/(\alpha_c + \alpha_d)$. **(c)** The perseveration parameter (ϕ) estimated from participant data using the Hybrid model (**red points**) and the PSL model (**blue points**). **Violet points** show the ϕ values recovered from fitting the Hybrid model to simulations of PSL agents, whose original generative parameters are shown in blue. The recovered violet points are systematically lower than the generative blue points, indicating underestimation. Points and error bars represent mean ± SEM across participants/simulations.*

Priors in the Fitting Procedure Can Bias Parameter Estimates

When fitting the hybrid model, we followed the standard Maximum a Posteriori (MAP) estimation method used in all studies we consider (see "Methods: Maximum a Posteriori estimation (MAP)") (2–4,26). A key feature of this method is its reliance on prior distributions for the parameters, which, if too stringent, can bias the resulting estimates, particularly in small datasets. For instance, following previous research, we used a narrow normal distribution (N(0,1)) as a prior for the perseveration parameter (ϕ). Given that perseveration is generally positive in our datasets, we suspected such a prior could produce a systematic underestimation (i.e., shrinkage) of this parameter. If a shrunk perseveration process is unable to accommodate for repetition rates in the data (to which the model was fitted), then spurious learning asymmetries can emerge. In turn, this can (partially) account for a confirmatory-learning/perseveration tradeoff observed in the previous section.

To test this possibility, we conducted a parameter recovery study. We simulated datasets from the Hybrid model (based on the empirical parameters of all experiments) and then fitted these datasets with the same model (see "Methods: Comparison of Model Fitting Procedures"). As suspected, the MAP fitting procedure systematically underestimated the true generative ϕ parameters (mean difference between recovered and generative parameter = -0.29, p<0.001; Fig. 3a-b, top left). Concurrently, the recovered confirmation bias was significantly overestimated for MAP fittings (mean difference = 0.031, p<0.001; Fig. 3a-b, bottom left).

We then considered the alternative of a Maximum Likelihood Estimation (MLE) parameter-estimation procedure, which relaxes assumptions about parameter priors (see "Methods: Maximum Likelihood Estimation (MLE)"). While this procedure still produced a significant underestimation of ϕ (mean difference = -0.13, p=0.007; Fig. 3a-b, top-right), it was nevertheless significantly smaller than that observed using the MAP procedure (p=0.001; Fig. 3b, top). Moreover, MLE led to a slight underestimation of a confirmation bias (mean difference = -0.013, p<0.001; MLE vs MAP, p<0.001; Fig. 3a-b, bottom right), indicating a reduced susceptibility to an artificial inflation of learning rate asymmetries.

Next, we tested whether these two fitting procedures produced different results for our datasets. When we fitted participants' data with a hybrid model the MAP procedure, across experiments, resulted in a lower estimate for the ϕ parameter (p<0.05; Fig. 3c right panel), but a significantly higher estimate for a confirmation bias (absolute, p<0.01; normalized, p<0.05) compared to the MLE procedure (see Fig. 3c left and middle panels). This same trade-off was observed when the hybrid model was fit to PSL simulations (based on PSL fitted parameters; Fig. 3d). However, even though MLE produced a reduction in the spurious confirmation bias (absolute, p<0.001; normalized, p<0.01), this spurious confirmation bias was still significantly positive (absolute, p=0.028; normalized, p=0.02; see SI 2.2 for equivalent of Figure 2 for MLE procedure). Notably, when we simulated a simple symmetric learning model without perseveration, the MLE procedure did not produce a consistent, spurious confirmation bias (see SI 2.1), unlike the MAP procedure (see SI 1.3). This indicates that MLE is less susceptible to estimation biases that go beyond the primary issue of model mimicry.

Our results suggest that applying shrinking priors to the perseveration parameter biases the parameter fits, leading to an underestimation of perseveration and a corresponding spurious inflation of confirmation bias. However, even non shrinking procedures like MLE can still yield biased learning rate asymmetries.

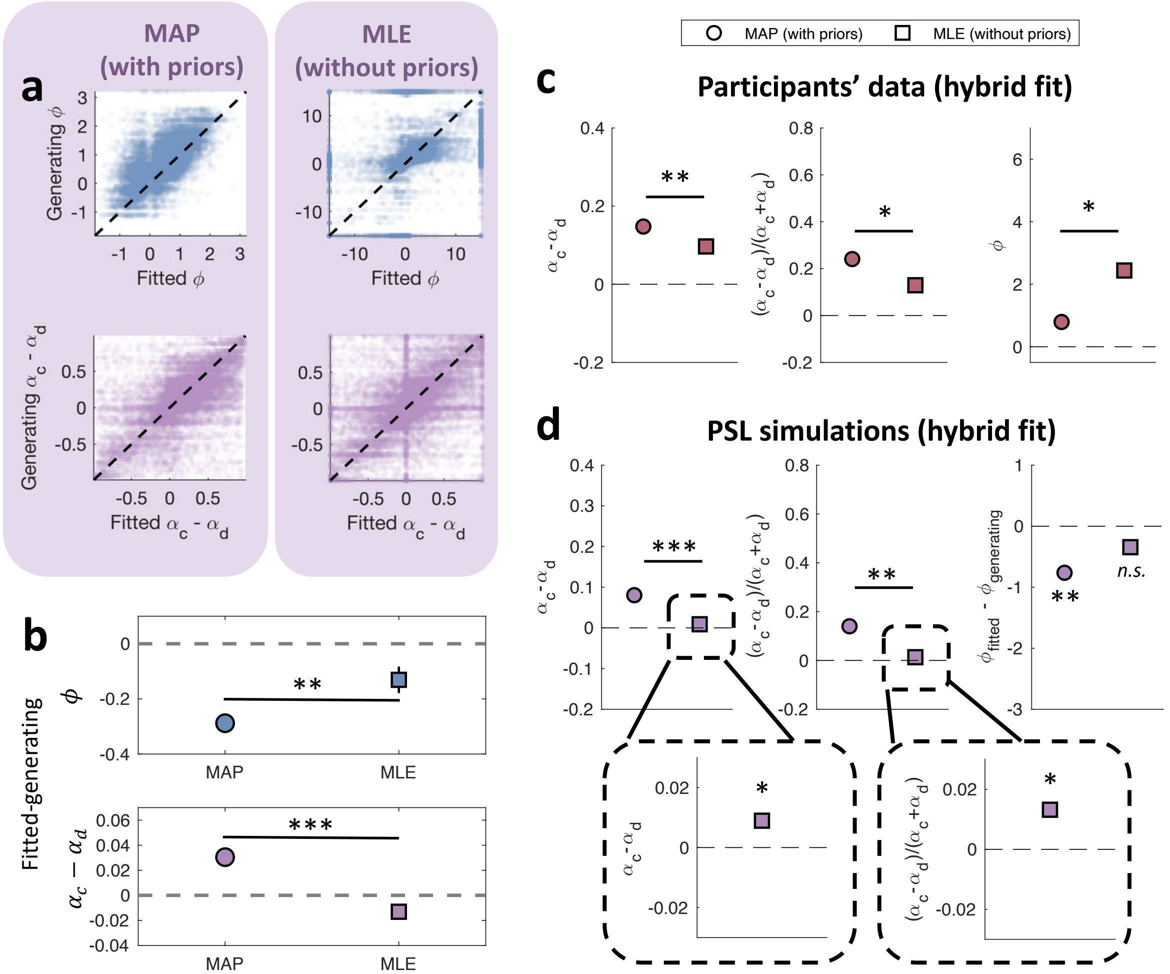


Figure 3: Priors in the fitting procedure bias parameter estimates and inflate spurious confirmation bias.(a) *Parameter recovery for parameters in the hybrid model. The scatter plots show the generative ϕ (top) and confirmation bias ($\alpha_c - \alpha_d$) (bottom) parameters used in the hybrid simulations (y-axes) versus the parameters recovered by the fitting procedure (x-axes), using either MAP (left) or MLE (right) estimation.* ***(b)*** *Parameter estimation errors for each procedure. This panel shows the difference between the fitted and generative parameters from (a). Negative values indicate that the fitting procedure underestimates the parameter, while positive values indicate overestimation.* ***(c)*** *Trade-off between perseveration and confirmation bias for empirical data. Average results across 10 experiments from fitting the hybrid model to participant data using either MAP (circles) or MLE (squares). Using the MAP procedure results in a inflated estimated confirmation bias (absolute, left; normalized, middle) and a lower estimated ϕ (right).* ***(d)*** *Similar to (c) but based on fits for the PSL agent simulations. Note that for the leftmost panel represents the underestimation of ϕ by each procedure (i.e., the differences between the ϕ fitted with the hybrid model, and the one used to generate the PSL simulation). Fitting the PSL datasets with MLE still generates a spurious confirmation bias (zoomed-in panels at the bottom). Circles/squares represent the mean result when fitting with MAP/MLE respectively, and error bars represent the standard error of the mean (s.e.m.) across the experimental means. (n.s. $p>0.05$; * $p < 0.05$; ** $p < 0.01$;*** $p < 0.001$).*

Asymptotic Dissociation of Perseveration and Confirmation Bias in Large Datasets

Our finding of a spurious confirmation bias using MLE (Fig. 3d) may be surprising at first glance. However, MLE guarantees unbiased estimation only at the "asymptotic limit" of very large datasets. Researchers collecting hundreds of trials per participant might assume they are within this asymptotic range, but is such an assumption justified? To address this question, we next tested whether, and at what rate, a spurious learning-asymmetry bias decreases as a function of set size. We simulated the behavior of PSL agents on the P2 task design (a common full-feedback design in the literature) using the average best-fit empirical parameters from that task. We systematically varied the size of the simulated datasets, from 1 to 100 sessions (where each session encompassed 4 blocks of 24 trials). For each size, we generated 4000 datasets and subsequently fitted them with the full hybrid model (using MLE) to assess the resulting parameter estimates (see "Methods: Asymptotic Parameter Estimation").

The analysis revealed that the magnitude of the spurious confirmation bias decreases rapidly as a function of set size (Fig. 4a). However, this artifact remained significant until a very large number of trials—around 7,000—were included in the simulation. In contrast, tasks in the datasets we consider here usually employ around 200-400 trials, and on this basis are highly susceptible to a spurious confirmation bias. Examining the confirmatory and disconfirmatory learning rates, separately, revealed that a spurious confirmation bias in smaller datasets was driven by a systematic underestimation of the confirmatory learning rate (α_c), which quickly converges to the true value (for tasks larger than 600 trials), coupled with an even greater underestimation of the disconfirmatory learning rate (α_d) which takes much longer to converge (only converging after 9,000 trials; Fig. 4b). The perseveration parameters were also misestimated in small datasets (see SI 2.2 for the asymptotic behavior of the perseveration parameters). This small-sample issues are substantially worse for MAP estimation, as overcoming the influence of parameter priors requires an unfeasibly large number of trials (~100,000) to cleanly dissociate perseveration and learning asymmetries (see SI 1.5).

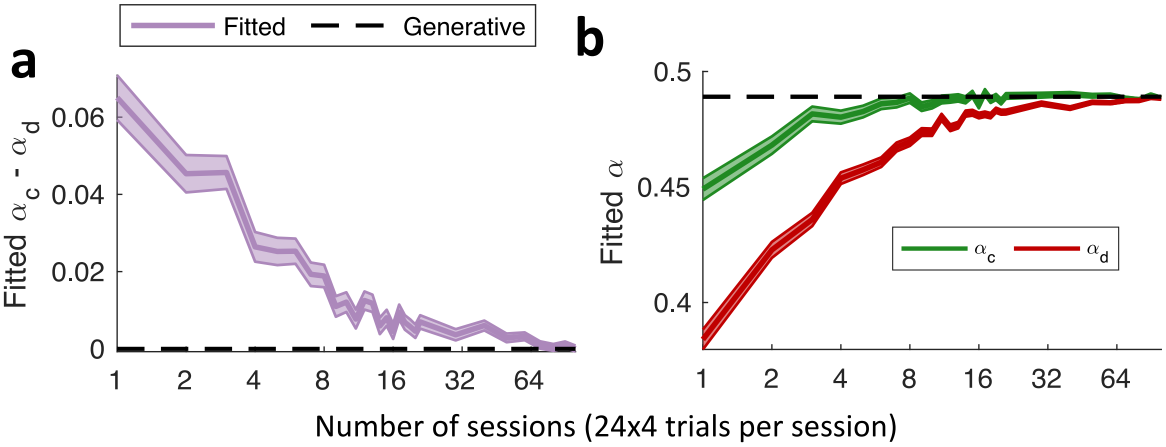


Figure 4: Spurious confirmation bias is reduced with dataset size. *The figure shows parameter estimates obtained by fitting the Hybrid model to simulations of PSL agents, as a function of the number of sessions in the simulated dataset (where each session encompasses 4 blocks of 24 trials) and the fitting procedure used. The simulations were based on the task in the second experiment (full feedback*

condition) from Palminteri et al. (2017). ***(a)*** *The absolute confirmation bias ($\alpha_c - \alpha_d$) fitted with MLE is large for small datasets and asymptotes towards zero for large datasets (around 7,000 trials).* ***(b)*** *Disaggregation of the effect in (a), showing the separately recovered confirmatory (α_c) and disconfirmatory (α_d) learning rates. In all panels, solid lines and shaded areas represent the mean ± SEM across 4000 simulations. Dashed lines represent the true generative parameter value used in the simulations. Note the difference x and y-axes scales used in left-side and right-side panels.*

Perseveration Level Influences the Magnitude of Spurious Confirmation Bias

So far, our results establish that perseveration can generate a spurious confirmation bias, but leaves open the question of how different levels of perseveration may contribute to this artifact. To better understand this, we conducted a parameter sweep analysis to characterize the relationship between the strength of an agent's perseveration and the magnitude of the resulting spurious bias.

To test this, we simulated perseverative symmetrically-learning (PSL) agents in experiment P2 (using the average empirical parameters from such experiment), and then we systematically varied the perseveration parameter (ϕ) between -7 and 7 while holding the other PSL parameters fixed. We then fitted these simulated datasets with the hybrid model using MLE (see “Methods: Parameter Sweep of the Perseveration Parameter”). The analysis revealed that an artifactual confirmation bias increased as a function of the perseveration parameter (Fig. 5a-b). However, this relationship was highly asymmetric: positive values of ϕ generated a strong spurious confirmation bias, while negative values of ϕ generated a much weaker spurious disconfirmation bias (see SI 2.5 for parameter sweep of other generating parameters).

This asymmetry has an important implication in that an individual with a negative ϕ (anti-perseverative) will typically exhibit a spurious disconfirmation bias, whilst a population of agents with a negative average ϕ can still produce a group level positive confirmation bias . This is because strong positive artifacts generated by perseverative individuals in the population can overwhelm weaker negative artifacts from the anti-perseverative individuals. To test this explicitly, we generated 2,000 simulations of a PSL population with an average negative ϕ (average ϕ = -1, p<0.001) (Fig. 5c) (see “Methods: Simulation of a Population with Negative Perseveration”). When this dataset was fitted with the hybrid model, we detected a significantly positive confirmation bias, for both the absolute (mean=0.04, p<0.001; Fig. 5d) and normalized metrics (mean=0.05, p<0.001; Fig. 5e). These results illustrate that a significant, albeit spurious, confirmation bias can emerge at the group level even when the population’s average tendency is anti-perseverative (see SI 1.6 for equivalent conclusions using MAP estimation).

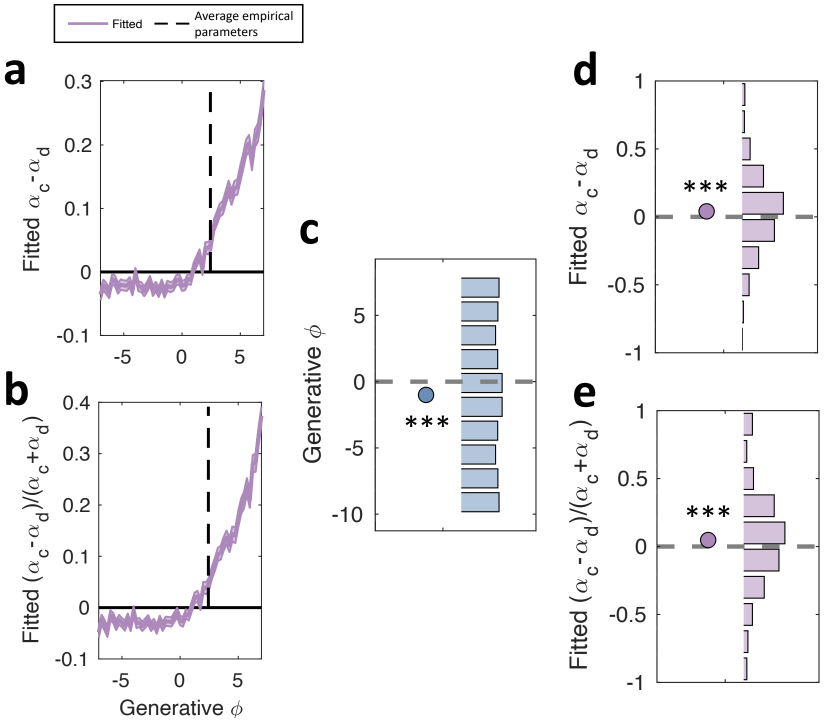


Figure 5: The perseveration parameter (ϕ) systematically influences the magnitude of spurious confirmation bias. (a) *Recovered absolute confirmation bias ($\alpha_c - \alpha_d$) as a function the generative ϕ parameter. The dashed vertical line represents the average empirical parameter in the experiment.* ***(b)*** *Same as (a), but for the normalized confirmation bias metric, $(\alpha_c - \alpha_d)/(\alpha_c + \alpha_d)$.* ***(c-e)*** *Results from a separate simulation where the generative ϕ parameter was drawn from a perfectly symmetric distribution centred at zero.* ***(c)*** *The distribution of the generative perseveration parameter (ϕ) from the PSL simulations, which follows a uniform distribution with a mean of -1.* ***(d)*** *The absolute confirmation bias detected when fitting these PSL simulations. The fitted bias is significantly positive (p<0.001).* ***(e)*** *Same as in (d), but for the normalized confirmation bias, which is also significantly positive (p<0.001).* In all panels, dots represent the mean parameter, and histograms represent the parameter distribution.

Testing Against a Null Distribution of Spurious Confirmation Bias

Given these problems of unbiased asymmetric-learning estimation using the hybrid model, we propose a new statistical test for the null hypothesis (H0) that individuals engage in symmetric learning (and might exhibit choice perseveration) against the alternative hypothesis (H1) of asymmetric learning + perseveration. For that, we generate a null distribution for the hybrid model's estimation of learning asymmetry under H0. Comparing the empirical confirmation bias against this null distribution allows us to test, even in “non asymptotic datasets”, whether it is statistically significant or simply reflects biased parameters.

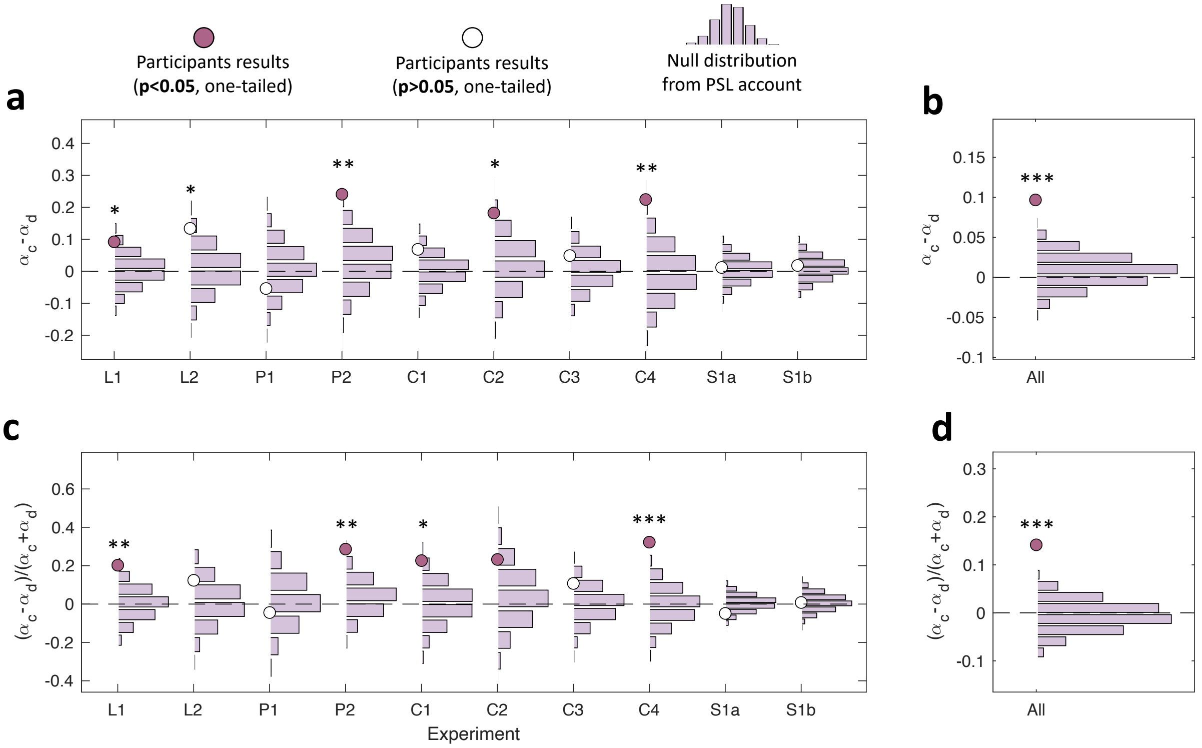
To generate this null distribution, we followed a precise parametric bootstrapping procedure. The process began by fitting each participant's data with the PSL model using MLE, to obtain their individual best-fit parameters. This allows us to estimate the extent of participants’ perseveration under H0. Using these parameters we then generated 1,001 simulated behavioral datasets of the

PSL model for each participant. Next, each of these simulated datasets was fitted with the full hybrid model, yielding a large pool of potential "spurious bias" (i.e., the difference between positive and negative learning rates) estimates for every participant. For an experiment with *N* participants, this resulted in a total pooled set of $N \times 1{,}001$ artifactual bias estimates. Finally, to create a null distribution for the mean population-level spurious-effect expected under H0, we used a bootstrap analysis on this pooled data. We generated 10,000 bootstrap samples; each sample was created by drawing *N* bias estimates at random without replacement from the entire pooled set and calculating their mean. This method effectively simulates drawing new samples of N participants from a population defined by the observed distribution of perseverative tendencies. This process created the final null distribution of 10,000 plausible mean artifactual biases against which the empirical results could be compared. We tested for a positivity/confirmation biased by calculating a p-value as the proportion of null samples (out of 10,000) which were larger or equal in magnitude to the empirical effect (group level difference between positive and negative learning rates).

Correcting for perseverative (spurious confirmation) artifacts this analysis provides evidence that a confirmation bias is less widespread than previously thought. Only half (i.e. 5 out of 10) analyzed experiments demonstrated a significant absolute confirmation bias ($p < 0.05$ for L1, L2, P2, C2 and C4; see Fig. 6a), while using a normalized metric, 4 out of 10 experiments showed a significant effect ($p < 0.05$ for L1, P2, C1 and C4; see Fig. 6c). The impact of this stricter hypothesis testing is best illustrated by contrasting its results with those obtained from a direct t-test of the MLE estimates (without controlling for spurious learning asymmetries). For instance, in Chambon et al. (C2), a direct fit produces a significant normalized confirmation bias. However, with our approach, we find this bias is not significantly greater than would be expected from a symmetric learning process coupled with choice perseveration (Fig. 6c). By contrast, the absolute confirmation bias in L2, which is non-significant when directly fitting the hybrid model, leads to a significant rejection of the null hypothesis under our new procedure (Fig. 6a). Importantly, our bootstrapping procedure generally produces fewer significant findings than the approach used in prior research of fitting a hybrid model with MAP estimation (Fig 6e), highlighting our method mitigates the risk for false detection of learning asymmetries above and beyond perseveration.

To examine the evidence for a confirmation bias across the entire dataset, using our novel test, we computed a bootstrapped meta-analytic null distribution by averaging across the single-study null distributions. This revealed a significant confirmation bias at the meta-analytic level, for both the absolute and normalized metrics ($p < 0.001$; see Fig. 6b,d), suggesting that the overall evidence supports the presence of genuine asymmetric value updating, even after rigorously accounting for estimation biases.

Finally, in a "sanity check" validation-study, we also applied our statistical procedure to synthetic data generated from the PSL model and confirmed that our method does not reject H0 in this case (see SI 2.4). The same approach failed when a bootstrapping test was based on the MAP, rather than ML, estimation methods (see SI 1.7).



e

	Absolute confirmation bias $[\alpha_c - \alpha_d]$											Normalized confirmation bias $[(\alpha_c - \alpha_d)/(\alpha_c + \alpha_d)]$										
	L1	L2	P1	P2	C1	C2	C3	C4	S1a	S1b	All	L1	L2	P1	P2	C1	C2	C3	C4	S1a	S1b	All
Original conclusion	✓	✓	✗	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓
New method conclusion	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	✓

Figure 6: Testing empirical confirmation bias observed in participant data against a null distribution of artefactual bias generated from PSL simulations. *All model fits used in this procedure are based on MLE.* ***(a)*** *Absolute confirmation bias ($\alpha_c - \alpha_d$) for each of the 10 experiments. Each violet histogram represents the null distribution of the mean artifactual bias for that experiment, generated via a bootstrap procedure from PSL simulations. The overlying dots represent the mean empirical confirmation bias.* ***(b)*** *Meta-analytic equivalent of panel (a), showing the mean absolute confirmation bias aggregated across all 10 experiments. The violet histogram is the meta-analytic null distribution, and the red dot is the observed meta-analytic mean.* ***(c)*** *Same per-experiment analysis as panel (a), but for the normalized confirmation bias metric, defined as as $(\alpha_c - \alpha_d)/(\alpha_c + \alpha_d)$.* ***(d)*** *Meta-analytic result for the normalized confirmation bias, corresponding to panel (c). Dots are colored red if this empirical mean was significantly larger than zero in a one-tailed t-test, and white otherwise. Stars indicates significance level when testing whether the observed empirical bias (dot) is greater than the artefactual bias predicted by the null distribution (* $p < 0.05$; ** $p < 0.01$;*** $p < 0.001$).* ***(e)*** *Comparison between the conclusions drawn from the original method (Hybrid model with MAP, top row) versus our new bootstrapping method (bottom row). "X" indicates that a significant confirmation bias was not detected ($p > .05$), while "✓" indicates a significant bias was detected. The tables show that our method leads to fewer significant findings for both absolute (left) and normalized (right) confirmation bias.*

Limitations of current behavioral signatures

In addition to computational modeling, the literature has proposed several behavioral signatures to support the presence of learning asymmetries. However, because these signatures were often developed without considering the influence of perseveration, it is crucial to re-evaluate whether they can reliably dissociate the two processes.

One approach, proposed by Palminteri and Lebreton (2022), identifies three behavioral patterns associated with confirmation bias (5) (see "Methods: Simulations for Palminteri & Lebreton (2022) Signatures"). To test whether these signatures reliably dissociate confirmation bias from perseveration, we simulated behavior from both a confirmation bias model (with no perseveration) and our PSL model. First, in a task with two equally rewarding random bandits (each with a 50% probability of reward and non-reward), a confirmation bias model (compared to symmetric learning) predicts a stronger preference for one option. This occurs because once an early-selected choice elicits a reward, it is difficult to "erase" the impression that it is better, as this initial reward is weighted more heavily than subsequent losses. Our simulations confirm this intuition, but also shows a PSL model produces a similar pattern (Fig. 7a, top row). Second, in a reversal task, a confirmation bias model predicts slower adaptation after the reward contingencies reverse (compared to a symmetric learning account with no perseveration) because negative feedback for the pre-reversal better choice-option, is underweighted. Again, our PSL-simulations showed that perseveration produces similar effects (Fig. 7a, middle row). On this basis we can conclude both signatures cannot conclusively implicate a learning-rate asymmetry. Finally, when choosing between a safe and a risky option with equal mean rewards, a confirmation bias model predicts (for high levels of confirmation bias) a preference for the risky option as wins are overweighted compared to loses. Here, the PSL model produced the opposite pattern—a reduced preference for the risky option (Fig. 7a, bottom row). Thus, a preference for the risky option can indeed implicate a positivity bias beyond and above gradual perseveration. However, we believe the practical utility of this signature is limited. For example, in our simulation a significant preference for the risky option emerged only for levels of confirmation bias higher than those typically observed (e.g., higher than our meta-analytic estimate of around 0.1) with lower levels of confirmation bias yielding a preference for the safe option (Fig. 7a, bottom row, middle panel). Furthermore, a co-occurring perseverative tendency, which favors the safe option, can counteract and mask a true underlying confirmation bias (see Fig.7a bottom right panel). We also note that, although not the topic of the current paper, a preference for the risky option could also be explained by alternative risk-seeking mechanisms.

As alternative regression-based signature has been proposed by Katahira (2018) (22) (see "Methods: Simulations for Katahira (2018) Regression Signature"). This involves regressing the probability of repeating on a current trial t+1 a choice from trials t and t-1 (this analysis is restricted to trials where the same option was chosen at t and t-1) on rewards from both these trials. Intuitively, the interaction term between the two past rewards captures how the influence of a past outcome is modulated by a subsequent one. A negative interaction is the key putative signature of confirmation bias because if the learning rate on trial t is higher when this trial's outcome is a reward (compared to non-reward) then this serves to weaken the influence of the outcome of trial t-1 on the current choice. Conversely, a positive interaction would signify a disconfirmation bias.

While the authors provided a general mathematical intuition for this signature, it was only validated in a specific reversal task design, and only for positive perseveration values. This leaves its generalizability to other designs and negative perseveration as an open question. Our analyses shows that even within this reversal-design, agents with negative perseveration also produce a negative interaction term that might be confused with a confirmation bias (Fig. 7b top). We note that most empirical studies are better powered to test behavioral signatures at the group level rather than the individual level. The results in Figure 7b (right) indicate that this interaction signature can be negative at the group level even when group-level perseveration is positive, as participants with "negative perseveration" contribute more to towards a negative signature (than participants with positive perseveration contribute to its positivity).

Furthermore, when we tested this signature in a classic, non-reversal bandit task, we found that all levels of perseveration, both positive and negative, produced the same negative interaction signature as a confirmation bias (Fig. 7b bottom). This indicates this regression-based signature does not fully dissociate learning asymmetries from perseveration and is highly sensitive to aspects of the task design (see SI 3 for results in other common task designs).

The upshot is that the behavioral signatures are subject to being confounded with perseveration-contributions, or are only reliable under unrealistically high values of confirmation bias, limiting their practical utility.

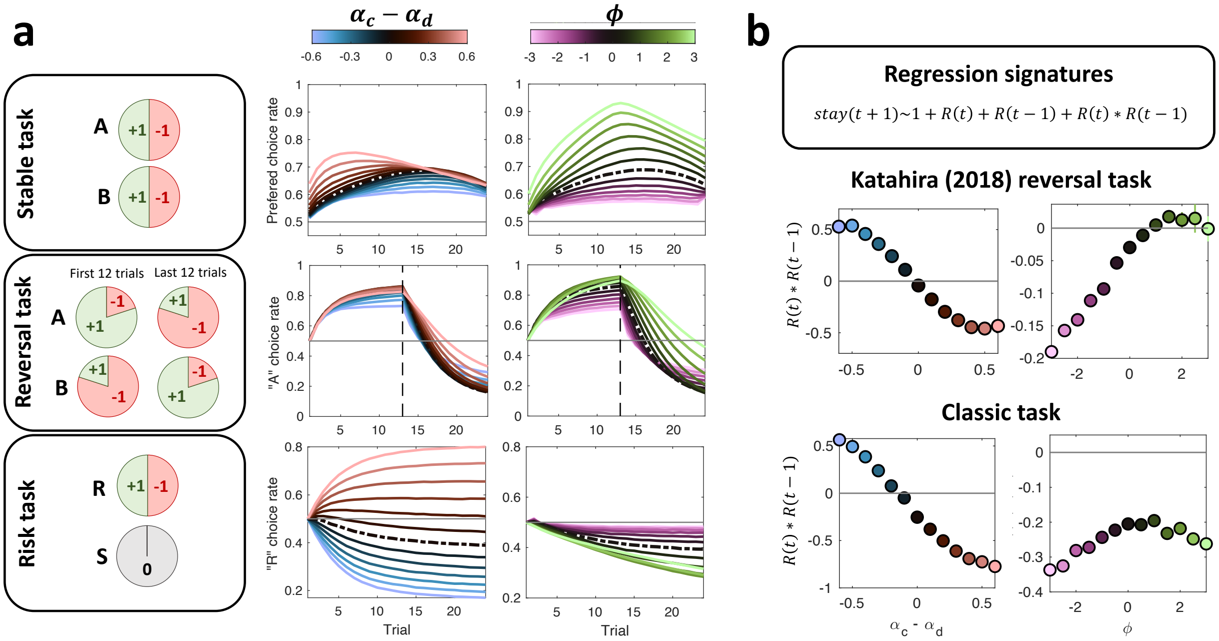


Figure 7: Behavioural patterns associated with confirmation bias (Palminteri & Lebreton, 2022).(a) *Behavioural signatures proposed in Palminteri and Lebreton 2022. For each task design, we simulated both a confirmation bias model (without perseveration) and our PSL models, sweeping over the parameters of interest ($\alpha_c - \alpha_d$ and ϕ respectively).* ***(top row)*** *Stable task signature. In a task with two random bandits (both with 50% reward probability), the development of a preference (y-axis, representing the choice rate for the most chosen bandit in a block) is more pronounced if participants exhibit both a confirmation bias or choice-perseveration.* ***(middle row)*** *Reversal task signature. In a task where the reward probabilities of the two bandits (80% and 20%) are flipped midway through the block, both*

confirmation bias and perseveration predict a reduced ability to adjust to the reversal. ***(bottom row)*** *Risk task signature. In a task where one option is safe (100% probability of getting 0 points), while the other is risky (50% probability of winning a point and 50% of losing a point), a higher confirmation bias predicts a higher preference for a risky option, while positive perseveration predicts an increased preference for the safe option. Dashed lines represent the choice rates for symmetric learning and no perseveration.* ***(b)*** *Regression signature proposed in Katahira 2018. This panel tests the robustness of a regression-based signature in a classic two-armed bandit task. The analysis uses a mixed-effects logistic regression model* $(stay(t+1) \sim R(t) * R(t-1))$*, restricted to consecutive choices of the same option.* ***(top panels)*** *Katahira's reversal task. We simulated a task with 70%/30% reward probabilities and 4 reversals. The left column shows that as confirmation bias increases, the interaction term becomes more negative. The right column shows that for positive values of perseveration (ϕ>0), as simulated in the original paper, we find no spurious interaction. However, for negative values of perseveration, a negative interaction emerges that can be confounded with confirmation bias.* ***(bottom panels)*** *We simulated a classic non-reversal bandit task with 25%/75% reward probabilities. The right column shows the interaction term for different confirmation bias levels. The right column shows that for all tested levels of perseveration we find a negative interaction term, demonstrating that the confirmation bias signature is not robust in this common task design.*

A proposal for a behavioural signature of a learning asymmetry

Given the limitations of extant behavioural signatures, we developed a new experimental design offering a model-agnostic signature that can dissociate confirmation bias from choice-perseveration (see “Methods: Simulations of New Behavioural Signature”). Our design features 4 bandits, whose outcomes are drawn from Gaussian distributions. These Gaussian distributions vary in their mean (high-mean, $\mu = 2$; low-mean, $\mu = 1$) and variance (high-variance, $\sigma = 1$; low-variance, $\sigma = 0.5$), in a 2x2 design (Fig. 8a). The task operates over two phases. In a learning-phase, bandits are presented across-trials in two pairs (low-mean/low-variance vs. high-mean/high-variance, and low-mean/high-variance vs. high-mean/low-variance), with participants receiving full feedback for both chosen and unchosen bandits (Fig. 8b, left). In this phase, high-mean bandits are mostly chosen and low-mean bandits mostly unchosen. If participants show a confirmation bias, then a high-mean bandit with higher variance will end up being perceived as better than the high-mean bandit with the lower variance (as the extreme positive values of the higher-variance bandit will be overweighted). Conversely, the low-variance low-mean bandit will be perceived as better than the high-variance low-mean bandit (whose extreme negative values will be overweighted). In a second generalization-phase, we offer a choice between two new pairings, where bandits of the same mean are paired against each other, without feedback, so as to prevent additional learning (Fig. 8b, right).

As explained, confirmation bias, during learning, predicts a preference, during generalization, for the high-variance bandit in the high-mean pair and for the low-variance bandit in the low-mean pair, while a disconfirmation bias predicts the opposite pattern (Fig. 8c-d, middle panels). Importantly, when learning is symmetric, this model predicts choice indifference. Furthermore, simulations of a PSL model also predicts no preference, since the paired bandits have the same expected values and have been sampled a similar number of times during first-phase learning (Fig. 8c-d, right panels). Therefore, unlike the previous signatures, this design creates a scenario where any significant deviation from chance performance in the generalization phase can be uniquely attributed to learning rate asymmetries, above and beyond perseveration. Another key advantage is that this signature also dissociates learning asymmetries from simple risk-seeking preferences. A risk-seeking agent would consistently prefer the high-variance option in both the high-mean and low-mean pairs. In contrast, a confirmation bias predicts a choice pattern that depends on the interaction between mean and variance: a preference for high-variance in the high-mean pair, but low-variance in the low-mean pair. This provides a unique and falsifiable prediction that distinguishes the learning asymmetry account from a general risk preference, a potential improvement over the "risk preference” signature (Fig. 7a, bottom) discussed in the previous section.

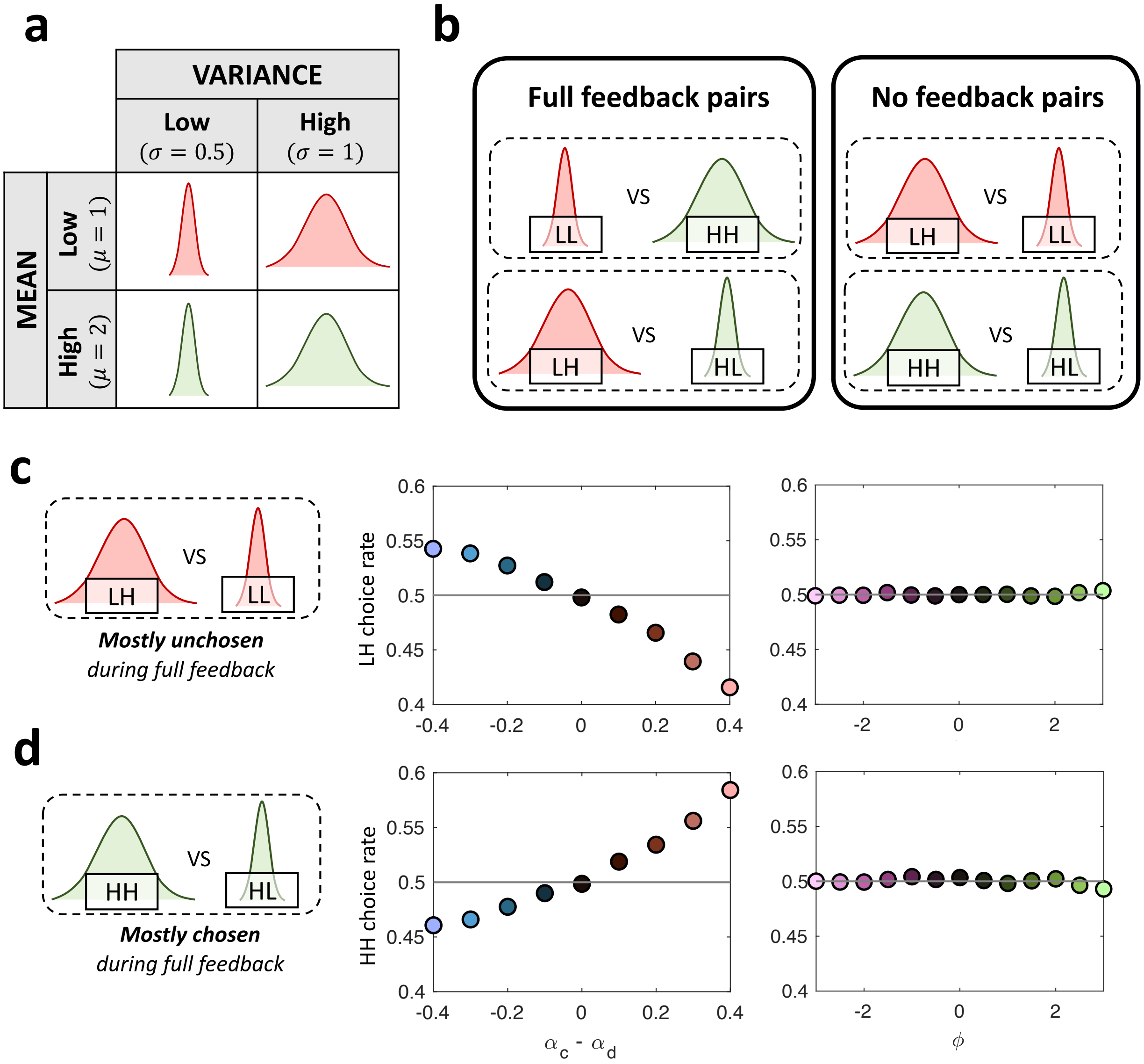


Figure 8: A novel behavioural signature to dissociate confirmation bias from choice perseveration. ***(a)*** *The task design involves four bandits whose outcomes are drawn from Gaussian distributions, varying in mean (high-mean, green; low-mean, red) and variance (high-variance, broad; low-variance, narrow).* ***(b)*** *The task has two phases. In the learning phase (left), participants learn about the bandits in two specific pairings (low-mean/low-variance vs. high-mean/high-variance, and low-mean/high-variance vs.*

high-mean/low-variance) with full feedback. In the generalization phase (right), bandits with the same mean are paired against each other, and choices are made without feedback. ***(c-d)*** *Predicted choice probabilities in the test phase. The confirmation bias model (middle panels) predicts a preference for the high-variance bandit in the low-mean pair and the high-variance bandit in the high-mean pair. Both the unbiased model and the PSL model (right panels) predict no preference (a choice rate of 0.5).*

DISCUSSION

Is there evidence for asymmetric learning in reinforcement learning that goes above and beyond a tendency to perseverate on previous choices (or seek novel ones)? Here we provide a critical re-evaluation of this question and introduce a novel approach to evaluating such evidence. We first demonstrate that previous approaches, using a hybrid model that includes both mechanisms acting simultaneously (26,27), can be biased in detecting a confirmation bias above and beyond perseveration. Indeed, using simulations we show that even when these methods are used, a perseverative symmetric-learning (PSL) agent is systematically misidentified as having systematic learning asymmetries, which are spurious by construction. This problem is amplified when using maximum a posteriori estimation (MAP) fitting procedures, common in much research, which apply shrinking priors over the perseveration parameter contributing to an inflated spurious confirmation bias. Although this is partially mitigated when using a non-shrinking maximum likelihood estimation (MLE) procedure, nevertheless a spurious bias is still detected.

Here we show this artifact is a consequence of partial mimicry between perseveration and learning-asymmetry, resulting in biased estimates of the latter due to a “small sample” problem, one that only disappears completely under unrealistic large number of trials. Additionally, since positive perseveration has a greater effect on learning asymmetries than negative perseveration, a spurious confirmation bias can emerge at the group level even in a population with overall anti-perseverative tendencies.

To address these issues, we propose a new statistical test for rejecting a null hypothesis of symmetric learning. This test still uses asymmetric learning estimates from the hybrid model but, critically, takes account of the possibility these estimates may be biased. Using a bootstrapping approach, to generate a null distribution of the spurious confirmation bias that would be expected under the assumption that individuals exhibit symmetric learning + perseveration, we can calculate a significance-level of empirical asymmetry estimates. Applying this method to a large dataset, we found that whilst evidence for confirmation bias is more nuanced than previously reported, it remains significant across several individual studies as well as at the meta-analytic level. Finally, we show that extant behavioral signatures for positivity/confirmation bias are also sensitive to perseveration, rendering them either invalid or of limited practical usefulness (5,22). Our broader aim is encouraging researchers to design novel tasks that are better suited for demonstrating unique contributions of learning-asymmetries, for which we propose one potential design.

Our findings invite a re-evaluation of the robustness of research linking individual differences to learning asymmetries. It has been proposed that these asymmetries might serve as a critical signature in a range of mental health disorders, including a reduced valence bias in major depression and anxiety (14,29,30), blunted learning from negative outcomes in OCD patients

(31), and increased learning from rewards in pathological gambling (31). Similar links have been drawn to developmental changes, such as learning rate asymmetries in adolescence (6), and even to political ideology, where conservatives are suggested to learn more from negative outcomes (32). However, our results urge caution in that studies that derive learning biases from choice behaviour risk misattributing effects of perseveration to asymmetric learning, a misattribution that fundamentally changes the interpretation of findings.

For example, consider a reported blunted learning from negative outcomes in OCD (31). Assume, however, that individuals with OCD exhibit greater choice perseveration. Our simulations show that when fitting behaviour with a hybrid model, a portion of this heightened perseveration would be misattributed to a spurious learning asymmetry. Consequently, even if the true learning process is unaffected by OCD, the model would incorrectly report an increased asymmetry in these patients. Furthermore, a parameter sweep analysis shows that the magnitude of the spurious learning-asymmetry can be systematically affected by other parameters, such as choice stochasticity, posing an additional challenge for interpreting between-subject correlations (see SI 1.7 and 2.5). Our findings, that the hybrid model yields biased estimated of asymmetric learning, highlight an urgent need for future research to develop tools that will allow for the estimation of learning asymmetries at the individual level in an unbiased way.

The broader implications of our findings extend beyond the specific confound between learning asymmetries and perseveration and speak to a general problem of model mimicry in reinforcement learning. The fundamental challenge in dissociating these processes is that they often make a common behavioral prediction, generating more choice repetition or switching than would be expected from a simple, symmetric-learning, agent devoid of perseveration. Importantly, several additional cognitive processes produce choice inertia and therefore may also be subject to model mimicry. For example, in partial feedback settings, 'fictive' updates to the unchosen option (in the opposite direction of the chosen one) can magnify value differences and increase choice repetition (33,34). Similarly, learning rates that adaptively decrease in stable environments can create choice 'stickiness', such that after learning it can take many negative outcomes to erase a preference for one bandit (35,36). Another potential source is stochasticity in the update process itself. Thus, if the learning process is 'noisy', a random positive fluctuation can locally inflate the value of a chosen option, rendering it more likely to be repeated and thus increasing choice hysteresis (37). It has also been proposed that the mere act of selecting an option increases its hedonic value, which could again contribute to choice hysteresis (38,39). We suspect at least some of these processes present similar problems of model-mimicry (with perseveration, learning asymmetries of among themselves) creating a complex web that needs careful scrutiny in future studies.

More broadly, our findings provide an important case-study and reminder for the research community as to how subtle assumptions about parameter priors (e.g., in MAP fitting) can cascade into spurious effects. Strikingly, we show that even ML estimates yield a spurious confirmation bias—raising cautionary flags for researchers who rely on this method. While it is often assumed that sufficiently large datasets locate them in an asymptotic (non-biased) estimation regime, our results show this is not necessarily the case, and attaining such a regime requires an unfeasibly large number of trials.

On a positive note, we suggest that the tools we have developed here can guide researchers' efforts to address issues of process-mimicry and biased estimates in reinforcement learning and in other domains of research. Our proposed bootstrapping method, for instance, can be adapted to any situation where parametric estimates are systematically biased. Closely related processes often interact to affect empirical phenomena and the current work is a step towards clarifying and validating their distinct contributions.

METHODS

Analysed datasets

We analysed the data from the same four studies as Palminteri (2022) (2–4,26), which encompass 10 two-armed bandit experiments. In the table below we detail how these studies differ in number of participants (N), number of trials (Trials), conditions (reward contingencies of the bandits and choice contingencies), and whether feedback was partial (only for the chosen option) or full (for both the unchosen and chosen options).

Table 1: Summary table of the studies analysed in this paper. *Unless otherwise specified, the number of trials was evenly distributed across conditions. Reward contingencies are represented as the percentage probability of reward for each bandit in a pair (separated by "/"). The studies by Chambon et al. (2020) also included different choice contingencies: 'free' choices operate in the traditional way; 'free+forced' interleaves free-choice trials with forced-choice trials where participants observe an external agent making a choice; in 'go' trials participants select a bandit by pressing a button, while in 'no-go' trials they select it by not pressing it.*

Study	Experiment	N	Trials	Number of conditions	Type of feedback
Lefevbre et al. (2017)	L1	50	96	4 reward contingencies *[25/25, 25/75, 25/75,75/75]*	Partial
	L2	35	96		Partial
Palminteri et al. (2017)	P1	20	192 *(96x2 sessions)*	4 reward contingencies *[50/50, 25/75,25/75, 17/83 (reversal)]*	Partial
	P2	20	192 *(96x2 sessions)*		Full
Chambon et al. (2020)	C1	24	720	2 reward contingencies *[60/90, 10/40]* x 2 choice contingencies [free, free + forced]**	Partial

	C2	24	640	2 reward contingencies *[60/90, 10/40]* x 1 choice contingency *[free + forced]*	Full
	C3	30	360	2 reward contingencies *[50/50, 30/70]* x 1 choice contingency *[free + forced]*	Partial
	C4	20	600	2 reward contingencies *[50/50, 30/70]* x 2 choice contingencies *[go, no-go]*	Partial
Sugawara and Katahira (2021)	S1a*	143	192 *(96x2 sessions)*	4 reward contingencies *[50/50, 25/75,25/75, 17/83 (reversal)]*	Partial
	S1b*	143	192 *(96x2 sessions)*		Full

*The two experiments were conducted in a blocked manner with the same participants.

** free+forced trials conditions contain twice as many trials as free conditions.

Computational models

We modeled participant behavior using a set of reinforcement learning models based on a standard Q-learning framework. In all models, the expected value (Q-value) of an option i is updated after receiving a reward R based on a prediction error, $\delta_t(i) = R_t - Q_t(i)$. The models differ in how they update values and how they translate those values into choices.

Confirmation Bias (CB) Model

This model tests for learning asymmetries in the absence of perseveration. It updates the value of the chosen option using two distinct learning rates: a confirmatory learning rate (α_c) for better-than-expected outcomes ($\delta_t > 0$) and a disconfirmatory learning rate (α_d) for worse-than-expected outcomes ($\delta_t < 0$).

$$Q_{t+1}(C) = \begin{cases} Q_t(C) + \alpha_c \cdot \delta_t(C)\ if\ \delta_t(C) > 0 \\ Q_t(C) + \alpha_d \cdot \delta_t(C)\ if\ \delta_t(C) < 0 \end{cases}$$

In partial-feedback tasks only the Q-value of the chosen option (C) is updated. However, in full-feedback tasks, where outcomes for the unchosen (U) are also provided, the Q-value of the unchosen option is also updated, using a reversed logic to the chosen option:

$$Q_{t+1}(U) = \begin{cases} Q_t(U) + \alpha_c \cdot \delta_t(U)\ if\ \delta_t(U) < 0 \\ Q_t(U) + \alpha_d \cdot \delta_t(U)\ if\ \delta_t(U) > 0 \end{cases}$$

Choices are made via a standard softmax function that only considers the learned Q-values, weighted by an inverse temperature parameter (β):

$$p(choose\ A\ over\ B) = \frac{1}{1 + e^{-\beta(Q_A - Q_B)}}$$

This model has 3 free parameters: $\{\alpha_c,\ \alpha_d,\ \beta\}$.

Perseverative Symmetrically Learning (PSL) Model

This model tests for choice perseveration in the absence of learning asymmetries. It updates Q-values using a single, symmetric learning rate (α) for all prediction errors, regardless of their sign.

$$Q_{t+1}(i) = Q_t(i) + \alpha \cdot \delta_t(i)$$

In addition to Q-values, this model includes a choice trace, $C_t(i)$, for each option, which is updated with an accumulation rate (τ):

$$C_{t+1}\ (i) = \begin{cases} C_t(i) + \tau \cdot \big(1 - C_t(i)\big)\ if\ i = C \\ C_t(i) + \tau \cdot \big(0 - C_t(i)\big)\ if\ i = U \end{cases}$$

Choices are made via a hybrid softmax function that is a weighted sum of the Q-values (scaled by β) and the choice traces (scaled by a perseveration parameter, ϕ):

$$p(choose\ A\ over\ B) = \frac{1}{1 + e^{-\beta(Q_A - Q_B)} + e^{-\phi(C_A - C_B)}}$$

This model has 4 free parameters: $\{\alpha,\ \beta,\ \tau,\ \phi\}$.

Hybrid Model

This model allows for both learning asymmetries and choice perseveration. It combines the features of the two previous models. Value updating is asymmetric, using separate learning rates for confirmatory (α_c) and disconfirmatory (α_d) outcomes, identical to the CB model. The choice rule is identical to the PSL model, incorporating both Q-values and choice traces. This model has 5 free parameters: $\{\alpha_c,\ \alpha_d,\ \beta,\ \tau,\ \phi\}$.

Model Versions for C1, C2 and C3

Some experiments in Chambon et al. (2020) (C1, C2, and C3) included "forced-choice" trials in which participants observed an external agent making a choice. Following the winning model from the original paper, these trials were modeled with specific update rules. The Q-value of the option chosen by the agent was updated using a single, separate learning rate parameter (α_{forced}), applied to both positive and negative prediction errors based on the outcome of the chosen option:

$$Q_{t+1}(C) = Q_t(C) + \ \alpha_{forced} \cdot \delta_t(C)$$

Additionally, C2 also included outcomes for the unchosen option in forced trials, whose Q-values were also updated using the same learning rate (α_{forced}):

$$Q_{t+1}(U) = Q_t(U) + \ \alpha_{forced} \cdot \delta_t(U)$$

On forced trials, the choice trace for both options was updated as if they were unchosen, meaning the trace for both simply decayed:

$$C_{t+1}\,(i) = C_t(i) + \tau \cdot \big(0 - C_t(i)\big)\; for\; both\; i = C\; and\; i = U$$

Parameter fitting procedures

Maximum Likelihood estimation (MLE)

In most of the study, we optimized model parameters (θ) using Maximum Likelihood Estimation (MLE) by minimizing the negative log-likelihood of the choice data given the model:

$$\hat{\theta}_{MLE} = argmin_{\theta}[-\log\,(P(Data|Model,\theta)]$$

To perform the minimization, we used the *fmincon* function in MATLAB. To ensure that the estimated parameters were psychologically plausible and to aid convergence, we set wide, upper and lower bounds for each parameter ($\beta \in [0{,}15]$, $\alpha \in [0{,}1]$, $\phi \in [-15{,}15]$). To avoid settling in local minima, the optimization procedure was run 5 times for each participant/simulation, using random starting points for the parameters uniformly sampled between the fitting bounds.

Maximum a Posteriori estimation (MAP)

As an alternative to MLE, we also estimated model parameters for each participant using Maximum a Posteriori (MAP) estimation, a Bayesian approach that incorporates prior beliefs about the parameters. We implemented this by minimizing the sum of the negative log-likelihood of the choice data given the model and the negative loglikelihood of the parameters given their prior probabilities:

$$\hat{\theta}_{MAP} = argmin_{\theta}[-\log\big(P(Data|Model,\theta)\big) - P(\theta)]$$

We used the same priors used in previous work by Palminteri (and re-used in Katahira’s work). For parameters bounded between [0,1] (α, τ), we used a beta distribution ($Beta(1.1, 1.1)$). For β we used a gamma distribution ($Gamma(1.2{,}5.0)$) and for ϕ we used a normal distribution ($N(0{,}1)$)

To perform the minimization, we again used the *fmincon* function in MATLAB. The optimization procedure was run 5 times for each participant/simulation, each with different random starting points randomly sampled from the same uniform distribution as in the MLE procedure.

Model simulations

Quantifying Spurious Learning Asymmetries (Figure 2)

To quantify the magnitude of spurious learning asymmetries that can emerge from PSL behavior, we conducted a simulation and re-fitting procedure. For each participant in each of the 10 experiments, we first took their individual best-fitting parameters as estimated from the PSL model. Using these parameters, we then generated 1,001 simulated behavioral datasets for that participant. Each simulation was run on the task structure (i.e., number of trials, reward contingencies) as the original experiment. This procedure resulted in a large set of simulated data generated by agents with symmetric learning but empirically-derived levels of perseveration. Finally, each of these simulated datasets was fitted with the Hybrid model to test for spurious confirmation bias (plotted as violet points in Fig. 2).

Comparison of Model Fitting Procedures (Figure 3)

To compare the MAP and MLE fitting procedures, we conducted two main analyses. First, to directly quantify the parameter recovery for the hybrid model under each procedure (Figure 3a-b), we performed the following steps. For each participant, we took their best-fitting hybrid model parameters as estimated by the MAP procedure and used them to simulate 40 new datasets per participant. We then fitted these simulated datasets with the hybrid model using the same MAP procedure and compared the recovered parameters to the original generative parameters. We repeated this entire process using the MLE procedure (i.e., generating data from MLE-fitted parameters and recovering with MLE).

Second, to assess how the choice of fitting procedure impacts the estimation of confirmation bias and perseveration in both empirical and simulated data (Figure 3c-d), we performed two further analyses. To examine the impact on empirical data, we fitted the original participant datasets with our hybrid model using both MAP and MLE and compared the resulting parameter estimates (Figure 3c). To examine the impact on the generation of spurious bias, we took the PSL model parameters previously estimated from the data (using both MAP and MLE) and used them to simulate new PSL datasets. We then fitted these simulated datasets with the hybrid model using the corresponding procedure (i.e., data generated based on MAP parameters was fitted with MAP) and compared the magnitude of the resulting spurious confirmation bias (Figure 3d).

Asymptotic Parameter Estimation (Figure 4)

To test whether the hybrid model fitted with MLE can dissociate learning asymmetries from perseveration in asymptotically large datasets, we conducted a simulation analysis (Figure 4a-b). We generated behavioural data from PSL agents using the average empirical ML estimated parameters from experiment P2 ($\beta = 1.79, \alpha = 0.49, \tau = 0.38, \phi = 2.43$), a task design with full feedback (see Table 1). We systematically varied the length of the simulated experiments, creating datasets containing between 1 and 90 sessions. Specifically, we simulated datasets with 1, 2, 3, and so on, up to 21 sessions. For larger datasets, we increased the length in steps of ten, simulating datasets with 30, 40, 50, and so on, up to 100 sessions. As in the original experiment, each session comprised 4 blocks of 24 trials. For each dataset length, we generated 4000 independent simulations. Finally, each simulated dataset was fitted with the Hybrid model. We then examined the recovered parameters to assess whether the spurious confirmation bias diminished and whether the generative perseveration parameters were accurately estimated as the number of trials increased.

Parameter Sweep of the Perseveration Parameter (Figure 5a-b)

The goal of this analysis was to map the relationship between the strength of an agent's perseverative tendency and the magnitude of the resulting artifactual confirmation bias.

We used the average empirical parameters $\{\alpha,\ \beta,\ \tau\}$ obtained from fitting the PSL model with MLE to experiment P2 as a base. We then systematically varied the perseveration parameter (ϕ), from -7 to 7 in steps of 0.25. For each of these ϕ values, we generated 1,001 PSL simulations on the P2 task structure. Finally, each simulated dataset was fitted with the Hybrid model using MLE, and we extracted the resulting absolute and normalized confirmation bias to produce the plots in Figure 5a-b.

Simulation of a Population with Negative Perseveration (Figure 5c-e)

This analysis was designed to explicitly test whether a spurious confirmation bias can emerge at the group level even when the population has a negative perseveration ($\phi < 0$).

We generated a set of 2,000 PSL simulations of the P2 task. The base parameters $\{\alpha,\ \beta,\ \tau\}$ were again taken from the average empirical MLE fits of experiment P2. The ϕ value for each simulation was drawn from a uniform distribution between -10 and 8 (Fig. 5c). We then fitted this entire dataset of 2,000 simulations with the Hybrid model using MLE and examined the distribution of the recovered absolute and normalized confirmation biases to test if their means were significantly different from zero (Fig. 5d-e).

Simulations for Palminteri & Lebreton (2022) Signatures (Figure 7a)

To test whether the behavioral patterns proposed by Palminteri and Lebreton (2022) could emerge from choice perseveration, we simulated behavior from both a CB model and a PSL model. For the CB model simulations, we used the average empirical β parameter obtained from experiment P2 (fitted with the CB model; $\beta = 2.21, \alpha_c = 0.52, \alpha_d = 0.15$), while the base learning rate was fixed at 0.4. We then systematically varied the confirmation bias from -0.6 to 0.6 in steps of 0.1. The confirmatory and disconfirmatory learning rates were derived as $\alpha_c = 0.4 + cb/2$ and $\alpha_d = 0.4 - cb/2$. For the PSL model simulations, we used the average empirical parameters $\{\alpha, \beta, \tau\}$ from experiment P2 (fitted with the PSL model; $\beta = 1.79, \alpha = 0.49, \tau = 0.38, \phi = 2.43$). We then systematically varied the perseveration parameter, ϕ, from -3 to 3 in steps of 0.5.

For both model types, we generated 50,000 simulations for each parameter value on three different single-block (24-trial) task designs. The Stable Task featured two bandits, both with a 50% reward probability. The Reversal Task featured two bandits with 80% and 20% reward probabilities, which reversed after trial 12. Finally, the Risk Task featured a safe (deterministic) option that always yielded 0 points, and a risky (probabilistic) option that yielded +1 or -1 points with 50% probability each.

Simulations for Katahira (2018) Regression Signature (Figure 7b)

To test the robustness of the regression-based signature proposed by Katahira (2018), we simulated behavior from the CB and PSL models on two different task designs. We used the same parameter sweep procedure described for Figure 6a. For each parameter level, we generated 1,000 datasets, each consisting of 20 simulated agents. We then applied a logistic regression analysis to each dataset to obtain a distribution of 1,000 interaction term coefficients. Specifically, we regressed the probability of staying with the same choice on trial $t + 1$ (a binary variable, $stay$, coded as 1 for stay, 0 for switch) based on the reward from the current trial ($R(t)$)

and the previous trial ($R(t-1)$), and their interaction ($stay \sim R(t) * R(t-1)$). This analysis is restricted to trials where the same option was chosen at t and t-1. Rewards were coded as 1 for a reward and 0 for a non-reward.

This procedure was conducted on two task designs. The first was the task simulated in Katahira's original paper, a single 200-trial block with two bandits (70%/30% reward probabilities) that reversed three times (at trials 51, 101, and 151). The second was a Classic Bandit Task, a single 50-trial block with two bandits (25%/75% reward probabilities) and no reversals.

Simulations of New Behavioural Signature (Figure 8)

To generate the predictions for our novel behavioural signature, we again simulated behaviour from the CB and PSL models, following the same parameter sweep procedure described for Figure 6a. For each parameter level, we generated 10,000 datasets. The task design featured four bandits whose outcomes were drawn from Gaussian distributions with varying means (high-mean: $\mu = 2$; low-mean: $\mu = 1$) and standard deviations (high-variance/risky: $\sigma = 1$; low-variance/safe: $\sigma = 0.5$). The task consisted of two phases. The initial learning phase featured 50 trials for each of two bandit pairs (low-mean/low-variance vs. high-mean/high-variance, and low-mean/high-variance vs. high-mean/low-variance) with full feedback. This was followed by a generalization phase with 50 no-feedback trials for each of two new pairs composed of bandits with the same mean value (high-mean/low-variance vs. high-mean/high-variance, and low-mean/low-variance vs. low-mean/high-variance). We then calculated the choice probabilities in this generalization phase to generate the behavioural signatures.

REFERENCES

1. Sutton RS, Barto AG. Reinforcement Learning: An Introduction.

2. Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S. Behavioural and neural characterization of optimistic reinforcement learning. Nat Hum Behav. 2017 Mar 20;1(4):1–9.

3. Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. PLOS Comput Biol. 2017 Aug 1;13(8):e1005684.

4. 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. Nat Hum Behav. 2020 Oct;4(10):1067–79.

5. Palminteri S, Lebreton M. The computational roots of positivity and confirmation biases in reinforcement learning. Trends Cogn Sci. 2022 Jul 1;26(7):607–21.

6. Rosenbaum GM, Grassie HL, Hartley CA. Valence biases in reinforcement learning shift across adolescence and modulate subsequent memory. Schlichting ML, Frank MJ, editors. eLife. 2022 Jan 24;11:e64620.

7. Sharot T. The optimism bias. Curr Biol. 2011 Dec;21(23):R941–5.

8. Sharot T, Korn CW, Dolan RJ. How unrealistic optimism is maintained in the face of reality. Nat Neurosci. 2011 Nov;14(11):1475–9.

9. Charness G, Dave C. Confirmation bias with motivated beliefs. Games Econ Behav. 2017 Jul 1;104:1–23.

10. Kappes A, Harvey AH, Lohrenz T, Montague PR, Sharot T. Confirmation bias in the utilization of others' opinion strength. Nat Neurosci. 2020 Jan;23(1):130–7.

11. Basol M, Roozenbeek J, van der Linden S. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. J Cogn. 3(1):2.

12. Rollwage M, Loosen A, Hauser TU, Moran R, Dolan RJ, Fleming SM. Confidence drives a neural confirmation bias. Nat Commun. 2020 May 26;11(1):2634.

13. Brown VM, Zhu L, Solway A, Wang JM, McCurry KL, King-Casas B, et al. Reinforcement Learning Disruptions in Individuals With Depression and Sensitivity to Symptom Change Following Cognitive Behavioral Therapy. JAMA Psychiatry. 2021 Oct;78(10):1–11.

14. Pike AC, Robinson OJ. Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. JAMA Psychiatry. 2022 Apr 1;79(4):313–22.

15. Vandendriessche H, Demmou A, Bavard S, Yadak J, Lemogne C, Mauras T, et al. Contextual influence of reinforcement learning performance of depression: evidence for a negativity bias? Psychol Med. 2023 Jul;53(10):4696–706.

16. Hahn U, Merdes C, Sydow M von. Knowledge through social networks: Accuracy, error, and polarisation. PLOS ONE. 2024 Jan 3;19(1):e0294815.

17. Lefebvre G, Deroy O, Bahrami B. The roots of polarization in the individual reward system. Proc R Soc B Biol Sci. 2024 Feb 28;291(2017):20232011.

18. Ciranka S, Linde-Domingo J, Padezhki I, Wicharz C, Wu CM, Spitzer B. Asymmetric reinforcement learning facilitates human inference of transitive relations. Nat Hum Behav. 2022 Apr;6(4):555–64.

19. Lowet AS, Zheng Q, Matias S, Drugowitsch J, Uchida N. Distributional Reinforcement Learning in the Brain. Trends Neurosci. 2020 Dec 1;43(12):980–97.

20. Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, et al. A distributional code for value in dopamine-based reinforcement learning. Nat 2020 5777792. 2020 Jan 15;577(7792):671–5.

21. Muller TH, Butler JL, Veselic S, Miranda B, Wallis JD, Dayan P, et al. Distributional reinforcement learning in prefrontal cortex. Nat Neurosci. 2024 Mar;27(3):403–8.

22. Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. J Math Psychol. 2018 Dec 1;87:31–45.

23. Miller KJ, Shenhav A, Ludvig EA. Habits without values. Psychol Rev. 2019 Mar;126(2):292–311.

24. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma. J Exp Psychol Gen. 2014 Dec;143(6):2074–81.

25. Xu HA, Modirshanechi A, Lehmann MP, Gerstner W, Herzog MH. Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. PLoS Comput Biol. 2021 Jun 3;17(6):e1009070.

26. Sugawara M, Katahira K. Dissociation between asymmetric value updating and perseverance in human reinforcement learning. Sci Rep. 2021 Feb 11;11(1):3574.

27. Palminteri S. Choice-Confirmation Bias and Gradual Perseveration in Human Reinforcement Learning. Behav Neurosci. 2022 Nov 18;137.

28. Toyama A, Katahira K, Kunisato Y. Examinations of Biases by Model Misspecification and Parameter Reliability of Reinforcement Learning Models. Comput Brain Behav. 2023 Dec 1;6(4):651–70.

29. Reinen JM, Whitton AE, Pizzagalli DA, Slifstein M, Abi-Dargham A, McGrath PJ, et al. Differential reinforcement learning responses to positive and negative information in unmedicated individuals with depression. Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol. 2021 Dec;53:89–100.

30. Jin Y, Gao Q, Wang Y, Dietz M, Xiao L, Cai Y, et al. Impaired social learning in patients with major depressive disorder revealed by a reinforcement learning model. Int J Clin Health Psychol IJCHP. 2023;23(4):100389.

31. Suzuki S, Zhang X, Dezfouli A, Braganza L, Fulcher BD, Parkes L, et al. Individuals with problem gambling and obsessive-compulsive disorder learn through distinct reinforcement mechanisms. PLOS Biol. 2023 Mar 14;21(3):e3002031.

32. Shook NJ, Fazio RH. Political ideology, exploration of novel stimuli, and attitude formation. J Exp Soc Psychol. 2009 Jul 1;45(4):995–8.

33. Ben-Artzi I, Kessler Y, Nicenboim B, Shahar N. Computational mechanisms underlying latent value updating of unchosen actions. Sci Adv. 2023 Oct 20;9(42):eadi2704.

34. Marciano-Romm D, Romm A, Bourgeois-Gironde S, Deouell LY. The Alternative Omen Effect: Illusory negative correlation between the outcomes of choice options. Cognition. 2016 Jan;146:324–38.

35. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci. 2007 Sep;10(9):1214–21.

36. Nassar MR, Wilson RC, Heasly B, Gold JI. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J Neurosci Off J Soc Neurosci. 2010 Sep 15;30(37):12366–78.

37. 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 Dec;22(12):2066–77.

38. Sharot T, Martino BD, Dolan RJ. How Choice Reveals and Shapes Expected Hedonic Outcome. J Neurosci. 2009 Mar 25;29(12):3760–5.

39. Sharot T, Fleming SM, Yu X, Koster R, Dolan RJ. Is Choice-Induced Preference Change Long Lasting? Psychol Sci. 2012 Oct 1;23(10):1123–9.