eLife

Neuroscience

Disinformation elicits learning biases

Juan Vidal-Perez ✉, **Raymond J Dolan**, **Rani Moran** ✉

Max Planck Centre for Computational Psychiatry and Ageing, University College London, London, United Kingdom • Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom • Department of Psychology, School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

Reviewed Preprint
v2 • September 4, 2025
Revised by authors

Reviewed Preprint
v1 • May 12, 2025

eLife Assessment

This study provides an **important** extension of credibility-based learning research with a well-controlled paradigm by showing how feedback reliability can distort reward-learning biases in a disinformation-like bandit task. The strength of evidence is **convincing** for the core effects reported (greater learning from credible feedback; robust computational accounts, parameter recovery) but **incomplete** for the specific claims about heightened positivity bias at low credibility, which depend on a single dataset, metric choices (absolute vs relative), and potential perseveration or cueing confounds. Limitations concerning external validity and task-induced cognitive load, and the use of relatively simple Bayesian comparators, suggest that incorporating richer active-inference/HGF benchmarks and designs that dissociate positivity bias from choice history would further strengthen this paper.

https://doi.org/10.7554/eLife.106073.2.sa0

Abstract

In open societies disinformation is often considered a threat to the very fabric of democracy. However, we know little about how disinformation exerts its impact, especially its influences on individual learning processes. Guided by the notion that disinformation exerts its pernicious effects by capitalizing on learning biases, we ask which aspects of learning from potential disinformation align with ideal “Bayesian” principles, and which exhibit biases deviating from these standards. To this end, we harnessed a reinforcement learning framework, offering computationally tractable models capable of estimating latent aspects of a learning process as well as identifying biases in learning. In two experiments, participants completed a two-armed bandit task, where they repeatedly chose between two lotteries and received outcome-feedback from sources of varying credibility, who occasionally disseminated disinformation by lying about true choice outcome (e.g., reporting non reward when a reward was truly earned or vice versa). Computational modelling indicated that learning increased in tandem with source credibility, consistent with ideal Bayesian principles. However, we also observed striking biases reflecting divergence from idealized Bayesian learning patterns. Notably, in one experiment individuals learned from sources that should have been ignored, as these were known to be fully unreliable. Additionally, the presence of disinformation elicited exaggerated learning from trustworthy information (akin to jumping to conclusions) and exacerbated a normalized measure of “positivity bias”

whereby individuals self-servingly boost their learning from positive, relative to negative, choice-feedback. Thus, in the face of disinformation we identify specific cognitive mechanisms underlying learning biases, with potential implications for societal strategies aimed at mitigating its harmful impacts.

Introduction

Disinformation is a pervasive and pernicious feature of the modern world (1). It is linked to negative social impacts that include public-health risks (2–4), political radicalization (5,6), violence (6—8) and adherence to conspiracy theories (8,9). Consequently, there is a growing interest in comprehending how false information propagates across social networks (10–12), including an interest in designing strategies to curb its impact (13–16) albeit with limited success to date (17). However, there is also a considerable knowledge lacuna regarding how individuals learn and update their beliefs when exposed to potential disinformation. Addressing this gap is crucial, as it has been suggested that disinformation propagates by exploiting cognitive biases (18–22). Thus, uncovering whether and how potential disinformation elicits distinct learning *biases* has the potential to better enable targeted interventions aimed at countering its harmful effects.

We start with an assessment of a prediction that individuals should modulate their learning as a function of the credibility of an information source, and learn more from credible, truthful, information-sources. This prediction is based on Bayesian principles of learning and on previous findings showing that individuals flexibly and adaptively adjust their learning rates in response to key statistical features of the environment. For example, learning is more rapid when observationuncertainty ("noise") decreases and in volatile, changing, compared to stable environments, particularly following detection of change-points that render re-change knowledge obsolete (23–25). Moreover, human choice is strongly influenced by social information of high (as opposed to low) credibility, such as majority opinions more confident judgments (26) and large group consensus (27). Additionally, people are disposed to follow trustworthy advisors (28), including those who have recommended optimal actions in the past (29,30).

We hypothesised that in a disinformation context individuals would show significant deviations from idealized Bayesian learning, reflecting a diversity of biases. First, filtering non-credible information is likely to be cognitively demanding (31), and this predicts such information would impact belief updating, even if individuals are aware it is untrustworthy. An additional consideration is that humans tend to learn more from positive self-confirming information (32–34), which presents one in a positive light. We conjectured, influenced by ideas from motivated-cognition (35), that low-credibility information provides a pathway for amplification of such a bias, as uncertainty regarding informationveracity might dispose individuals to self-servingly interpret positive information as true and explain-away negative information as false. A final additional consideration is the question of how exposure to potential disinformation impacts on learning from trusted sources. One possibility is that disinformation serves as a background context against which credible information would appear more salient. Alternatively, it might lead individuals to strategically reduce their overall learning in disinformation-rich environments, resulting in diminished learning from credible sources.

To address these questions, we adopt a novel approach within the disinformation literature by exploiting a Reinforcement Learning (RL) *experimental* framework (36). While RL has guided disinformation research in recent years (37–41), our approach is novel in using one of its most popular tasks: the "bandit task". This has the advantage that it provides computationally tractable models, that enable estimation of latent aspects of learning processes, such as belief updating. Moreover, our approach also enables an examination of the dynamics of belief updates over short timescales reflecting real-life engagements with disinformation, such as deciding

whether to share a post on social media. Moreover, bandit tasks in RL have proven success in characterizing key decision-making biases (e.g., positivity bias (42–44)), albeit in scenarios where learners receive accurate information. . Finally, a previous literature has suggested a role for reinforcement in the dissemination of disinformation, where individuals may receive positive reinforcement (likes, shares) for spreading sensationalized or misleading information on social media platforms, inadvertently reinforcing such behaviours and contributing to a disinformation proliferation (15,40,45).

We developed a novel "disinformation" version of the classical two-armed bandit task to test the effects of potential disinformation on learning. In the *traditional* two-armed bandit task (36,42,46), participants choose repeatedly between two unfamiliar bandits (i.e., slot machines), that provided rewards with different probabilities, to learn which bandit is more rewarding. Critically, in our *disinformation*-variant, true choice outcomes (reward or non-reward) were *latent*, i.e., unobservable. Instead, participants were informed about choice-outcomes by computer-programmed "feedback agents", who were disposed to occasionally disseminate disinformation by lying (reporting a reward when the true outcome was non-reward or vice versa). As these feedback-agents varied in truthfulness, this allowed us to test the effects of source-credibility on learning. We show across two studies that the extent of belief-updates increases as a function of source-credibility. However, there were striking deviations from ideal-Bayesian learning, where we identify several sources of bias related to processing potential disinformation. In one experiment, individuals learned from noncredible information that should in principle be ignored. Additionally, in both experiments, participants exhibited increased learning from trustworthy information when it was preceded by noncredible information and an amplified normalized positivity bias for non-credible sources, where individuals preferably learn from positive compared to negative feedback (relative to the overall extent of learning).

Results

Disinformation two-armed bandit task

We conducted a discovery (n=104) and main study (n=204). In both studies the learning tasks had the same basic structure but with a few subtle differences between them (see Discovery study and SI Discovery study methods). To anticipate, the results of both studies support mostly similar conclusions, and, in the results section, we focus on the main study, with the final results section detailing similarities and differences in findings across the two studies.

In the main study, participants (n=204) completed the *disinformation* two-armed bandit task. In the *traditional* two-armed bandit task (36,42,46), participants choose between two slot-machines (i.e., bandits) differing in their reward probability. Participants are not instructed about bandit rewardprobabilities but instead they are provided with veridical choice feedback (e.g., reward or nonreward), allowing participants to learn which bandit is more rewarding. By contrast, in our disinformation version *true* choice-outcomes were latent (i.e., unobserved) and participants were informed about these outcomes via three computerized feedback-agents, who had privileged access to the true outcomes.

Before commencing the task, participants were instructed that feedback agents could disseminate disinformation, meaning that they were disposed to lie on a random minority of trials, reporting a reward when the true outcome was a non-reward, or vice versa (**Fig. 1a**). Participants were explicitly instructed about the credibility of each agent (i.e., based on the proportion of truth-telling trials), indicated by a "star system": the 3-star agent was always truthful, the 2-star agents told the truth on 75% of the trials while the 1-star agent did so on 50% of the trials (**Fig. 1b**). Note that while the 1-star agent's feedback was statistically equivalent to random feedback, participants were not explicitly instructed about this equivalence. Each experimental block

encompassed 3 bandit pairs, each presented over 15 trials in a randomly interleaved manner. The agent on each trial was random subject to the constraint that each agent provided feedback for 5 trials for each bandit pair. Thus, in every trial, participants were presented with one of the bandit pairs and the feedback agent associated with that trial. Upon selecting a bandit, they then received feedback from the agent (**Fig. 1c**). Importantly, at the end of the experiment participants received a performance-based bonus based on *true* bandit outcomes, which could differ from agent-provided feedback. Within each bandit-pair one bandit provided a (true) reward on 75% of the trials and the other on 25% of trials. Choice accuracy, i.e., the probability of selecting the more rewarding bandit (within each pair), was significantly above chance (mean accuracy = 0.62, $t(203) = 19.94$, $p < .001$) and improved as a function of increasing experience with each bandit-pair (average overall improvement over 15 trials = 0.22, $t(203)=19.95$, $p<0.001$) (**Fig. 1d**).

Credible feedback promotes greater learning

A hallmark of RL value-learning is that participants are more likely to repeat a choice following positive compared to negative reward-feedback (henceforth, "feedback effect on choice repetition"). We tested a hypothesis, based on Bayesian reasoning, that this tendency would increase as a function of agent-credibility (**Fig. 3a**). Thus, in a binomial mixed-effects model we regressed choice-repetition (i.e., whether participants repeated their choice from the most recent trial featuring the same bandit pair; 0-switch; 1-repeat) on feedback-valence (negative or positive) and agent-credibility (1,2, or 3-star), where these are taken from the last trial featuring the same bandit pair (Methods for modelspecification). Feedback valence exerted a positive effect on choice-repetition ($b=0.72$, $F(1,2436)=1369.6$, $p<0.001$) and interacted with agent-credibility ($F(2,2436)=307.11$, $p<0.001$), with a feedback effect being greater for more credible agents (3-star vs. 2-star: $b= 0.91$, $F(1,2436)=351.17$; 3-star vs. 1-star: $b=1.15$, $t(2436)=24.02$; and 2-star vs. 1-star: $b=0.24$, $t(2436)=5.34$, all p's<0.001). Additionally, we found a positive effect of feedback for the 3-star agent ($b=1.41$, $F(1,2436)=1470.2$, $p<0.001$), and a smaller effect of feedback for the 2-star agent ($b=0.49$,$F(1,2436)=230.0$, $p<0.001$). These results support our hypothesis that learning increases as a function of information credibility (note that the feedback effect for the 1-star agent is examined below; see "Non-credible feedback elicits learning").

To confirm that increased learning based on information credibility is expected under an assumption that subjects adhere to Bayesian reasoning, we formulated two Bayesian models whereby the latent value of each bandit is represented as a distribution over the probability that a bandit is truly rewarding (**Fig. 2a, top panel**; Fig. S5c for an illustration of the model; for full model descriptions, see Methods). In the *instructed-credibility Bayesian* model, belief-updates are based on the *instructed* credibility of feedback-sources. This model is based on an idealized assumptions that during the feedback stage of each trial, the value of the chosen bandit is updated (based on feedback valence and credibility) according to Bayes rule reflecting perfect adherence to the instructed task structure (i.e., how true outcomes and feedback are generated). In contrast, a *free-credibility Bayesian* model, allows for the possibility that Bayes-rule updates during feedback are based on "distorted probabilities" (47), attributing *non-instructed* degrees of credibility to sources of false information (despite our explicit instructions on the credibility of different agents). In this variant, we fixed the credibility of the 3-star agent to 1 and estimated the credibility of 2 and 1-star agents as free parameters (which were highly recoverable; see Methods and SI 3.3). Both models additionally assumed uninformative, uniform, priors over reward probabilities of novel bandits and that learning is non-forgetful. Simulations based on both Bayesian models (see Methods) predicted increased learning as a function of feedback credibility (**Fig. 3b**; **top panels**; SI 3.1.1.1 Tables S3 and S4 for statistical analysis).

Next, we formulated a family of *non-Bayesian* computational RL models. Importantly, these models can flexibly express non-Bayesian learning patterns and, as we show in following sections, can serve to identify learning biases deviating from an idealized Bayesian strategy. Here, an assumption is that during feedback, the choice propensity for the chosen bandit (which here is

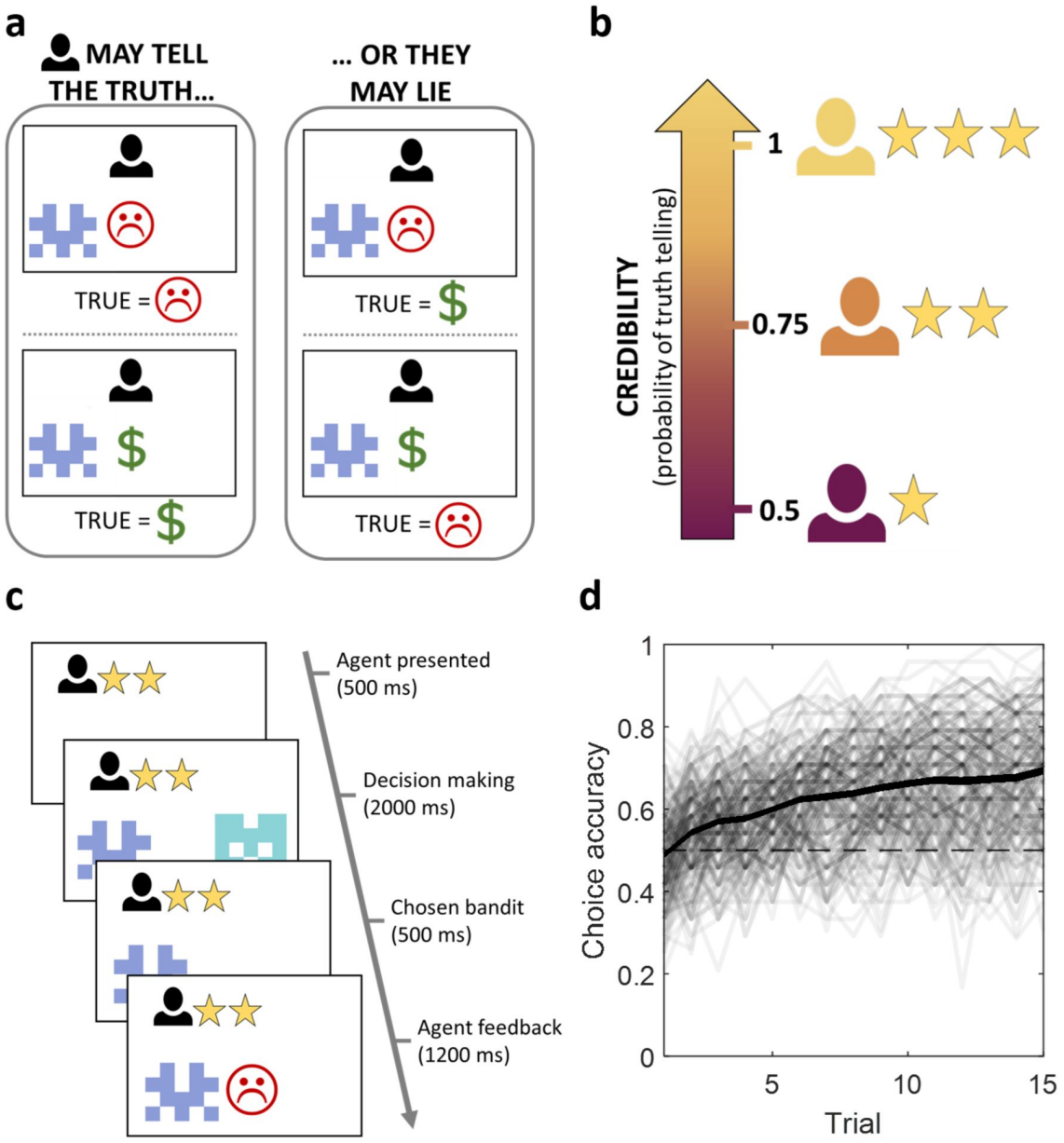

Figure 1

Task design and performance.

a, Illustration of agent-feedback. Each selected bandit generated a *true* outcome, either a reward or a non-reward. Participants *did not* see this true outcome but instead were informed about it via a computerised feedback agent (reward: dollar sign; non-reward: sad emoji). Agents told the truth on most trials (left panel). However, on a random minority of trials they lied, reporting a reward when the true outcome was a non-reward or vice versa (right panel). **b,** Participants received feedback from 3 distinct feedback agents of variable credibility (i.e., truth-telling probability). Credibility was represented using a starbased system: a 3-star agent always reported the truth (and never lied), a 2-star agent reported the truth on 75% of trials (lying on the remaining 25%), and a 1-star agent reported the truth half of the time (lying on the other half). Participants were explicitly instructed and quizzed about the credibility of each agent prior to the task. **c,** Trial-structure: On each trial participants were first presented with the feedback agent for that trial (here, the 2-star agent) and next offered a choice between a pair of bandits (represented by identicons) (for 2sec). Next, choice-feedback was provided by the agent. **d,** Learning curves. Average choice accuracy as a function of trial number (within a bandit-pair). Thin lines: individual participants; thick line: group mean with thickness representing the group standard error of the mean for each trial.

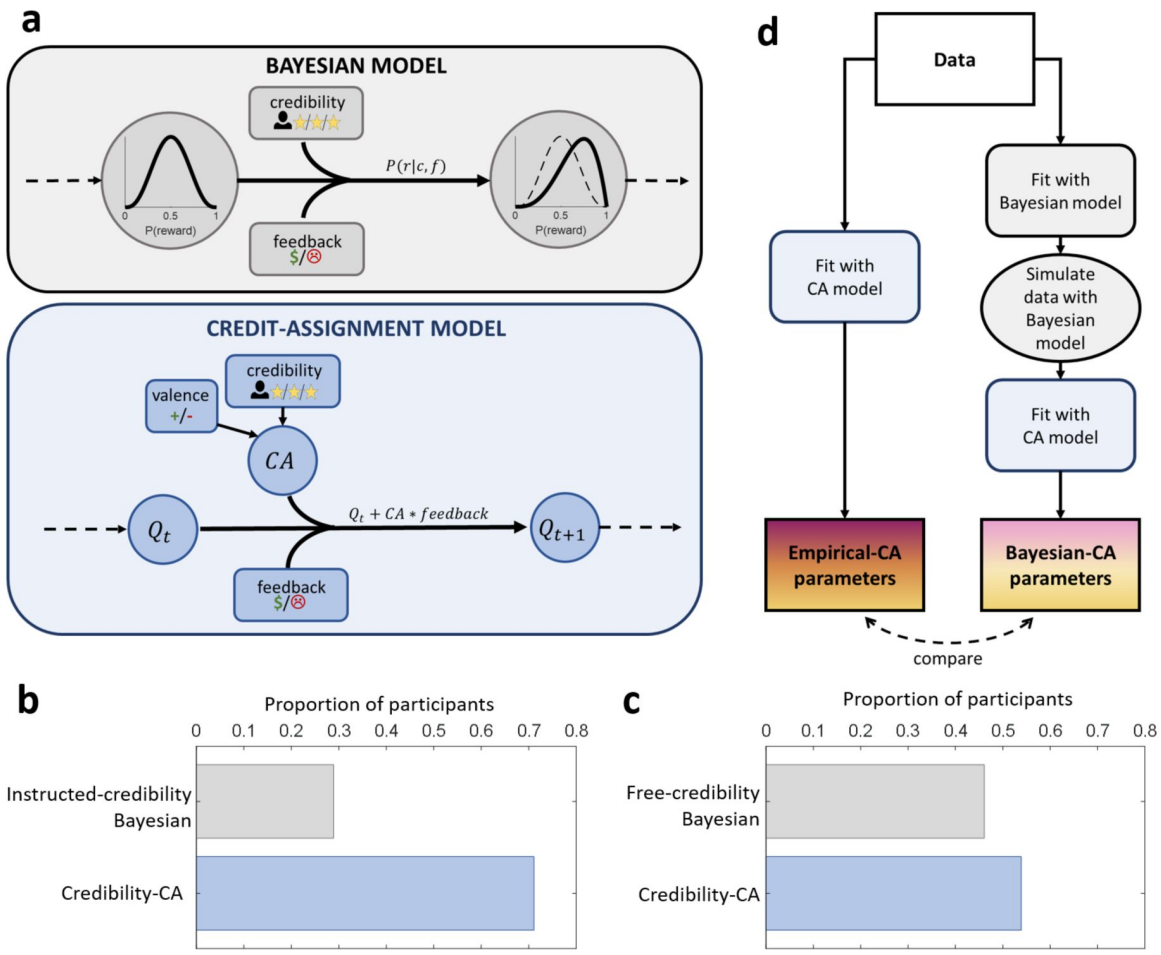

Figure 2

Computational models and cross-fitting method.

a, Summary of the two model families. In our Bayesian models (top panel), the observer maintains a belief-distribution over the probability a bandit is *truly* rewarding (denoted *r*). On each trial, this distribution is updated for the selected bandit according to Bayes rule, based on the valence (i.e., rewarding/non-rewarding; denoted *f*) and credibility of the trial's reward feedback (denoted *c*). In credit-assignment models (bottom panel), the observer maintains a subjective point-value (denoted *Q*) reflecting a choice propensity for each bandit. On each trial the propensity of the chosen bandit is updated based on a free CA parameter, quantifying the extent of value increase/decrease following positive/negative feedback. CA parameters can be modulated by the valence and credibility of feedback. **b,c,** Model selection between the credibility-CA model (without perseveration) and the two variants of Bayesian models. Most participants were best fitted by a credibility-CA model, compared to the instructed-credibility Bayesian model (b) or free-credibility Bayesian (c) models. **d,** Cross-fitting method: Firstly, we fit a Bayesian model to empirical data, to estimate its (ML) parameters. This yields the Bayesian learning token that comes closest to accounting for a participant's choices. Secondly, we simulate synthetic data based on the Bayesian model, using its ML parameters to obtain instances of how a Bayesian learner would behave in our task. Thirdly, we fit these synthetic data with a CA model, thus estimating "Bayesian CA parameters", i.e., CA parameters capturing the performance of a Bayesian model. Finally, we fit the CA model directly to empirical data to obtain "empirical CA parameters". A comparison of Bayesian and empirical CA parameters, allows us to identify, which aspects of behaviour are consistent with our Bayesian models, as well as characterize biases in behaviour that deviate from **our** Bayesian learning models.

represented by a point estimate, "Q value", rather than a distribution) either increases or decreases (for positive or negative feedback, respectively) according to a magnitude quantified by the free "Credit-Assignment (CA)" model parameters (48):

$$Q(chosen) \leftarrow \left(1 - f_Q\right) * Q(chosen) + CA(agent, valence) * F$$

where F is the feedback received from the agents (coded as 1 for reward feedback and −1 for nonreward feedback), while f_Q ($\in$[0,1]) is the free parameter representing the forgetting rate of the Q-value (**Fig. 2a, bottom panel**; Fig. S5b; see "Methods: RL models"). The probability to choose a bandit (say A over B) in this family of models is a logistic function of the contrast choice-propensities between these two bandits. One interpretation of this model is as a "sophisticated" logistic regression, where the CA parameters take the role of "regression coefficients" corresponding to the change in log odds of repeating the just-taken action in future trials based on the feedback (+/- CA for positive or negative feedback, respectively; the model also includes gradual perseveration which allows for constant logodd changes that are not affected by choice feedback) . The forgetting rate captures the extent to which the effect of each trial on future choices diminishes with time. The Q-values are thus exponentially decaying sums of logistic choice propensities based on the types of feedback a bandit received.

Within this model-family, different model variants varied as to how task-variables influenced CA parameters with the "null" model attributing the same CA to all feedback-agents (regardless of their credibility, i.e., a single free CA-parameter), whereas the "credibility-CA" model availed of three separate CA parameters, one for each feedback agent, thereby allowing us to test how learning was modulated by feedback-credibility. Using a bootstrap generalized-likelihood ratio test for modelcomparison (Methods) we rejected the null model (group level: p<0.001), in favour of the credibility-CA model. Furthermore, model-simulations based on participants best-fitting parameters (Methods) falsified the null model as it failed to predict credibility-modulated learning, showing instead, equal learning from all feedback sources (**Fig, 3b**; **bottom-left panel**). In contrast, the credibility-CA model successfully predicted increased learning as a function of credibility (**Fig. 3b**, **bottom-right panel**) (see SI 3.1.1.1 Tables S5 and S6).

After confirming CA parameters are highly recoverable (see Methods and SI 3.4), we examined how the Maximum Likelihood (ML) CA parameters from the credibility-CA model differed as a function of feedback credibility (**Fig. 3c**; see SI 3.3.1 for detailed ML parameter results). Using a mixed effects model (Methods), we regressed the CA parameters on their associated agents, finding that CA differed across the agents (F(2,609)=212.65, p<0.001), increasing as a function of agent-credibility (3-star vs. 2-star: b= 1.02, F(1,609)=253.73 ; 3-star vs. 1-star: b=1.24, t(609)=19.31; and 2-star vs. 1-star: b=0.22, t(609)=3.38, all p's<0.001).

Substantial deviations from our Bayesian learning models

We next implemented a model comparison between each of our Bayesian models and the credibility-CA model, using a parametric bootstrap cross-fitting method (Methods). We found that the credibility-CA model provided a superior fit for 71% of participants (sign test; p<0.001) when compared to the instructed-credibility Bayesian model, **Fig. 2b**; and for 53.9% (p=0.29) when compared to the free-credibility Bayesian model, **Fig 2c**). We considered using AIC and BIC, which apply "off-the shelf" penalties for model-complexity. However, these methods do not adapt to features like finite sample size (relying instead on asymptotic assumption) or temporal dependence (as is common in reinforcement learning experiments). In contrast, the parametric bootstrap cross-fitting method replaces these fixed penalties with empirical, data-driven criteria for model-selection. Indeed, modelrecovery simulations confirmed that whereas AIC and BIC were heavily biased in favour of the Bayesian models, the bootstrap method provided excellent model-recovery (See Fig. S20).

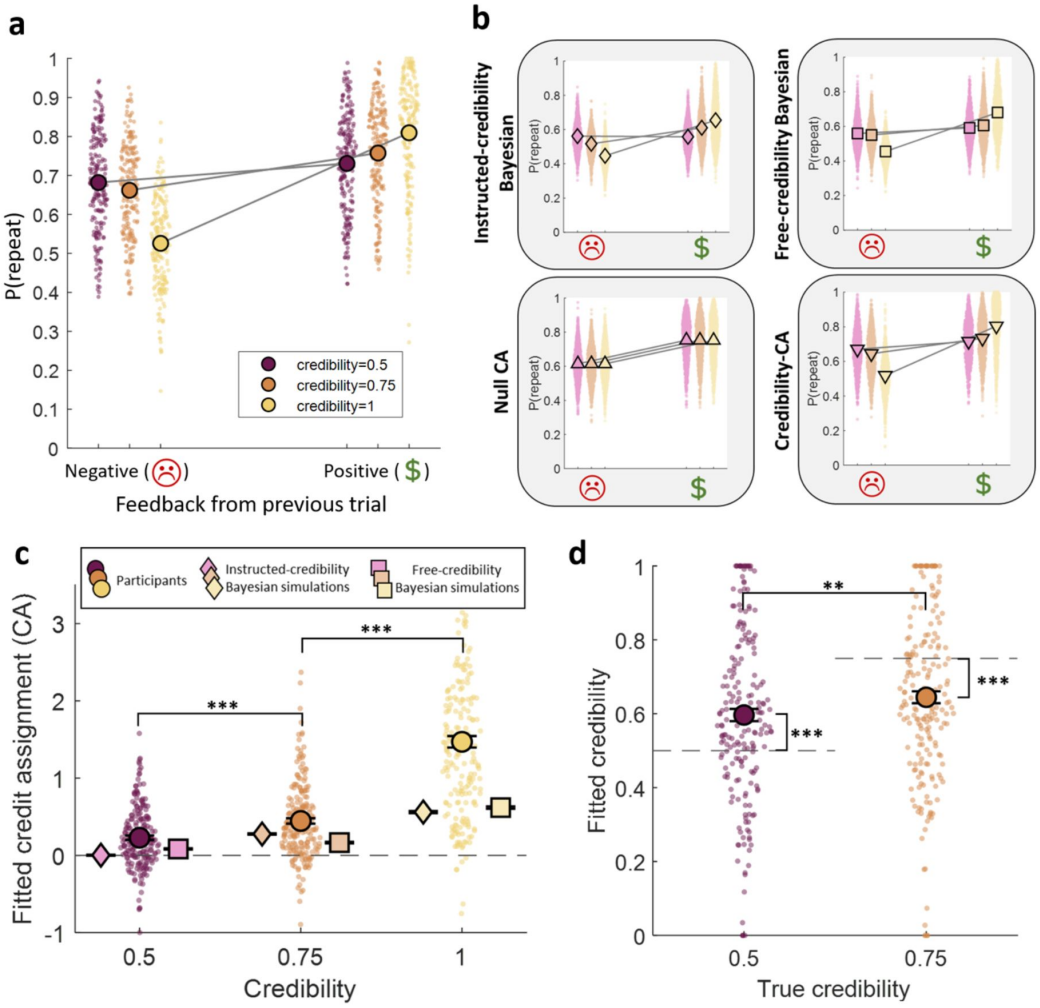

Figure 3

Learning adaptations to credibility.

a, Probability of repeating a choice as a function of feedbackvalence and agent-credibility on the previous trial for the same bandit pair. The effect of feedback-valence on repetition increases as the feedback credibility increases, indicating that more credible feedback has a greater effect on behaviour. **b,** Similar analysis as in panel a, but for synthetic data obtained by simulating the main models. Simulations were computed using the ML parameters of participants for each model. The null model (**bottom left**) attributes a single CA to all credibility-levels, hence feedback exerts a constant effect on repetition (independently of its credibility). The credibility-CA model (**bottom-right**) allowed credit assignment to change as a function of source credibility, predicting varying effects of feedback with different credibility levels. The instructed-credibility Bayesian model (**top left**) updated beliefs based on the true credibility of the feedback, and therefore predicted an increase effect of feedback on repetition as credibility increased. Finally, the free-credibility Bayesian model (**top right**) allowed for a possibility that participants use distorted credibilities for 1- star and 2-star agents when following a Bayesian strategy, also predicting an increase in the effect of feedback as credibility increased. **c,** ML credit assignment parameters for the credibility-CA model. Participants show a CA increase as a function of agent-credibility, as predicted by Bayesian-CA parameters for both the instructed-credibility and free-credibility Bayesian models. Moreover, participants showed a positive CA for the 1-star agent (which essentially provides random feedback), which is only predicted by cross-fitting parameters for the free-credibility Bayesian model. **d,** ML credibility parameters for a free-credibility Bayesian model attributing credibility 1 to the 3-star agent but estimating credibility for the two lying agents as free parameters. Small dots represent results for individual participants/simulations, big circles represent the group mean (a,b,d) or median (c) of participants' behaviour. Results of the synthetic model simulations are represented by diamonds (instructed-credibility Bayesian model), squares (free-credibility Bayesian model), upward-pointing triangles (null-CA model) and downward-pointing triangles (credibility-CA model). Error bars show the standard error of the mean. (*) p<.05, (**) p<0.01, (***) p<.001.

To further characterise deviations between behaviour and our Bayesian learning models, we used a "cross-fitting" method. Treating CA parameters as data-features of interest (i.e., feedback dependent changes in choice propensity), our goal was to examine if and how empirical features differ from features extracted from simulations of our Bayesian learning models. Towards that goal, we simulated synthetic data based on *Bayesian* agents (using participants' best fitting parameters), but fitted these data using the CA-models, obtaining what we term "Bayesian-CA parameters" (**Fig. 2d**; Methods). A comparison of these Bayesian-CA parameters, with empirical-CA parameters obtained by fitting CA models to empirical data, allowed us to uncover patterns consistent with, or deviating from, ideal-Bayesian value-based inference. Under the sophisticated logistic-regression interpretation of the CA-model family the cross-fitting method comprises a comparison between empirical regression coefficients (i.e., empirical CA parameters) and regression coefficient based on simulations of Bayesian models (Bayesian CA parameters). Using this approach, we found that both the instructed-credibility and free-credibility Bayesian models predicted increased Bayesian-CA parameters as a function of agent credibility (**Fig. 3c**; see SI 3.1.1.2 Tables S8 and S9). However, an in-depth comparison between Bayesian and empirical CA parameters revealed discrepancies from ideal Bayesian learning, which we describe in the following sections.

Non-credible feedback elicits learning

While our task instructions framed the 1-star agent as highly deceptive, lying 50% of the time, its feedback is statistically equivalent to entirely non-informative i.e., *random* feedback. Thus, participants should ignore and filter-out such feedback from their belief updates. Indeed, for the 1-star agent, simulations based on the instructed-credibility Bayesian model provided no evidence for either a positive effect of feedback on choice-repetition (mixed effects model described above; b=- 0.01, t(2436)=-0.41, p=0.68; **Fig 3b top-left**) or a positive Bayesian-CA (b=-0.01, t(609)=-0.31, p=0.76; **Fig. 3c**). However, contrary to this, we hypothesized that participants would struggle to entirely disregard non-credible feedback. Indeed, we found a positive effect of feedback on choice-repetition for the 1-star agent (mixed effects model, delta(M)=0.049, b=0.25, t(2436)=8.05, p<0.001), indicating participants are more likely to repeat a bandit selection after receiving positive feedback from this agent (**Fig. 3a**). Similarly, the CA parameter for the 1-star agent in the credibility-CA model was positive (b=0.23, t(609)=4.54, p<0.001) (**Fig. 3c**). The upshot of this empirical finding is that participants updated their beliefs based on random feedback (see Fig. S7 for analysis showing that this resulted in decreased accuracy rates).

A potential explanation for this finding is that participants *do* rely on a Bayesian strategy but "distort probabilities", attributing non-instructed degrees of credibility to lying sources (despite our explicit instructions on the credibility of different agents). Consistent with this, the ML-estimated credibility of the 1-star agent (**Fig. 3d**) was significantly greater than 0.5 (Wilcoxon signed-rank test, median=0.08, z=5.50, p<0.001), allowing the free-credibility Bayesian model to predict a positive feedback effect on choice-repetition (mixed-effects model: b=0.12, t(2436)=9.48, p<0.001; **Fig 3b topright**) and a positive Bayesian-CA (b=0.08, t(609)=3.32, p<0.001; **Fig. 3c**) for the 1-star agent. In our Discussion we elaborate on why it might be difficult to filter out this feedback even if one can explicitly infer its randomness.

Increased learning from fully credible feedback when it follows non-informative feedback

A comparison of empirical and Bayesian credit-assignment parameters revealed a further deviation from ideal Bayesian learning: participants showed an exaggerated credit-assignment for the 3-star agent compared with Bayesian models [Wilcoxon signed-rank test, instructed-credibility Bayesian model (median difference=0.74, z=11.14); free-credibility Bayesian model (median difference=0.62, z=10.71), all p's<0.001] (**Fig. 3a**). One explanation for enhanced learning for the 3-star agents is a contrast effect, whereby credible information looms larger against a backdrop of non-credible information. To test this hypothesis, we examined whether the impact of feedback

from the 3-star agent is modulated by the credibility of the agent in the trial immediately preceding it. More specifically, we reasoned that the impact of a 3-star agent would be amplified by a "low credibility context" (i.e., when it is preceded by a low credibility trial). In a binomial mixed effects model, we regressed choice-repetition on feedback valence from the last trial featuring the same bandit pair (i.e., the learning trial) and the feedback agent on the trial immediately preceding that last trial (i.e., the contextual credibility; see Methods for model-specification). This analysis included only learning trials featuring the 3-star agent, and context trials featuring the same bandit pair as the learning trial (**Fig. 4a**). We found that feedback valence interacted with contextual credibility ($F(2,2086)=11.47$, $p<0.001$) such that the feedback-effect (from the 3-star agent) decreased as a function of the preceding context-credibility (3-star context vs. 2-star context: $b= -0.29$, $F(1,2086)=4.06$, $p=0.044$; 2-star context vs. 1-star context: $b=-0.41$, $t(2086)=-2.94$, $p=0.003$; and 3-star context vs. 1-star context: $b=-0.69$, $t(2086)=-4.74$, $p<0.001$) (**Fig. 4b**). This contrast effect was not predicted by simulations of our main models of interest (**Fig. 4c**). No effect was found when focussing on contextual trials featuring a bandit pair different than the one in the learning trial (see SI 3.5). Thus, these results support an interpretation that credible feedback exerts a greater impact on participants' learning when it follows non-credible feedback in the same learning context.

Positivity bias in learning and credibility

Previous research has shown that reinforcement learning is characterized by a positivity bias, wherein subjects systematically learn more from positive than from negative feedback (42,44). One account is that this bias might result from motivated cognition influences on learning, whereby participants favour positive feedback that reflects well on their choices. We conjectured that feedback of ambiguous veracity (i.e., from the 1-star and 2-star agents) would promote this bias by allowing participants to explain-away negative feedback as a case of an agent-lying, while choosing to believe positive feedback. Following previous research, we quantified positivity bias in 2 ways: 1) as the *absolute* difference between credit-assignment based on positive or negative feedback, and 2) as the same difference but *relative* to the overall extent of learning. We note that the second, relative, definition, is more akin to "percentage change" measurements providing a control for the overall lower levels of credit-assignment for less credible agent. To investigate this bias across different levels of feedback credibility we formulated a more detailed variant of the CA model. To quantify the extent of a chosen-bandit's value increase or decrease - following positive or negative feedback respectively - the "credibility-valence-CA" variant included separate CA parameters for positive (CA+) and negative (CA-) feedback for each feedback agent. In effect, this model variant enabled us to test whether different levels of feedback credibility elicited a positivity bias (i.e., CA+ > CA-). Using a bootstrap generalized-likelihood ratio test for model comparison (Methods), we rejected, in favour of the valence-credibility-CA model, the null-CA model, the credibility-CA model and a "constant feedbackvalence bias" CA model, which attributed a common valence bias (CA+ minus CA-) to all agents (all group level: all p's<0.001). This test supported our choice of flexible CA parametrization as a factorial function of agent and feedback-valence.

After confirming the parameters of this model were highly recoverable (see Methods and SI 3.4), we used a mixed effects model to regress the ML parameters (**Fig. 5a**; see SI 3.3.1 for detailed ML parameter results) on their associated agent-credibility and valence (see Methods). This revealed participants attributed a greater CA to positive feedback than to negative feedback ($b=0.64$, $F(1,1218)=37.39$, $p<0.001$). Strikingly, for lying agents, participants selectively assigned credit based on positive feedback (1-star: $b=0.61$, $F(1,1218)=22.81$, $p<0.001$; 2-star: $b=0.85$, $F(1,1218)=43.5$, $p<0.001$), with no evidence for significant credit-assignment based on negative feedback (1-star: $b=- 0.03$, $F(1,1218)=0.07$, $p=0.79$; 2-star: $b=0.14$, $F(1,1218)=1.28$, $p=0.25$). Only for the 3-star agent, creditassignment was positive for both positive ($b=1.83$, $F(1,1218)=203.1$, $p<0.001$)

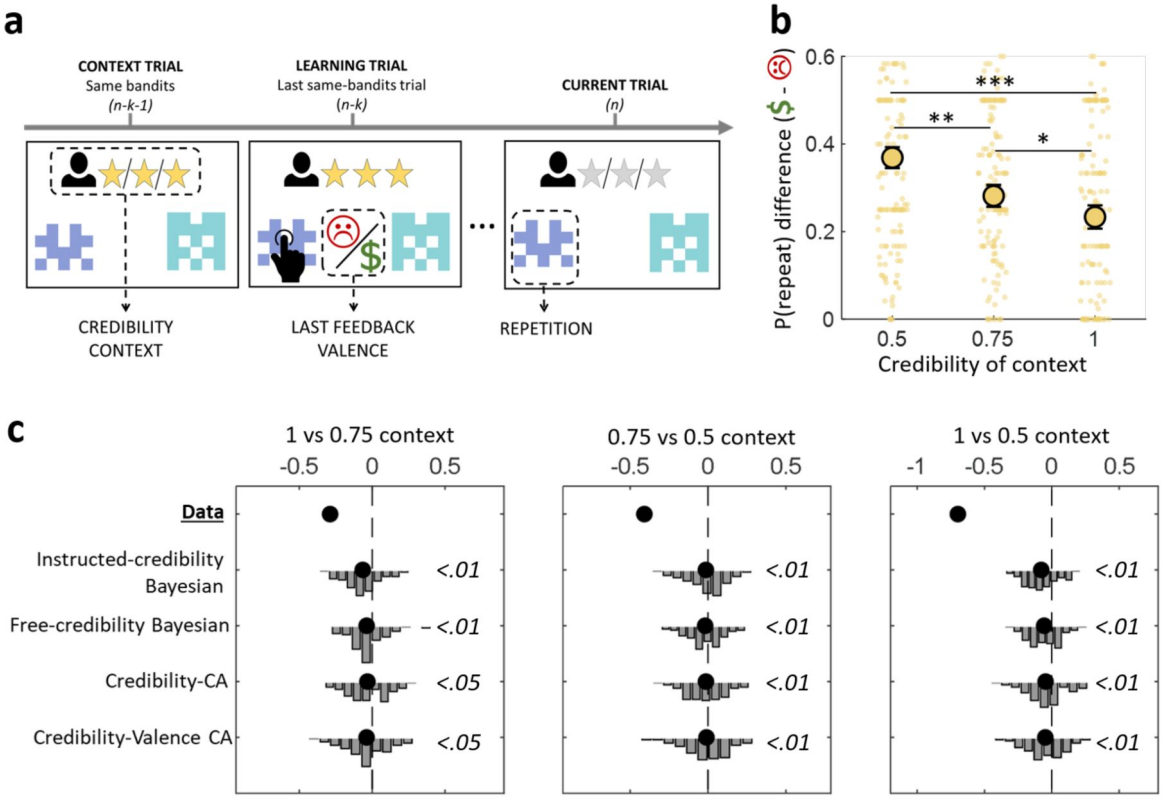

Figure 4

Contextual effects and learning.

a, Trials contributing to the analysis of effects of credibility-context on learning from the fully credible agent. We included only "current trials (n)" for which: 1) the last trial (trial nk) offering the same bandit pair (i.e., the learning trial) was associated with the 3-star agent, and 2) the immediately preceding context trial (n-k-1) featured the same bandit pair. We examined how choice-repetition (from n-k to n) was modulated by feedback valence on the learning trial, and by the feedback agent on the context trial. Note the greyed-out star-rating on the current trial indicates the identity of the current agent and was not included in the analysis. **b,** Difference in probability of repeating a choice after receiving positive vs negative feedback (i.e., feedback effect) from the 3-star agent, as a function of the credibility context. The 3-star agent feedback-effect is greater when preceded by a lower-credibility context, compared to a higher credibility context. Big circles represent the group mean, and error bars show the standard error of the mean. (*) p<.05, (**) p<0.01. **c,** We ran the same mixed effects model (regressing choice repetition of learning-trial feedback valence and on contextual credibility) on simulated data (See Methods: Model-agnostic analysis of contextual credibility effects on choice-repetition). The panels show contrasts in feedback-effect (from the 3-star agent in the learning trial) on choice-repetition between contextual credibility agent-pairs. None of our models predicted the contrast effects observed in participants. Histograms represent the distribution of regression coefficients based on 101 group-level synthetic datasets, simulated based on each model. The label right to each histograms represents the proportion of simulated datasets that predict an equal or stronger effect than the one observed in participants.

and negative (b=1.25, F(1,1218)=95.7, p=<0.001) feedback. We found no significant interaction effect between feedback valence and credibility on CA (F(2,1218)=0.12, p=0.88; **Fig. 5a-b**). Thus, there was no evidence for our hypothesis when positivity-bias was measured in absolute terms.

However, we found evidence for agent-based modulation of positivity bias when this bias was measured in relative terms. Here we calculated, for each participant and agent, a relative Valence Bias Index (rVBI) as the difference between the Credit Assignment for positive feedback (CA^+) and negative feedback (CA^-), relative to the overall magnitude of CA (i.e., $|CA^+| + |CA^-|$) (**Fig. 5c**). Using a mixed effects model, we regressed rVBIs on their associated credibility (see Methods), revealing a relative positivity bias for all credibility levels [overall rVBI (b=0.32, F(1,609)=68.16), 50% credibility (b=0.39, t(609)=8.00), 75% credibility (b=0.41, F(1,609)=73.48) and 100% credibility (b=0.17, F(1,609)=12.62), all p's<0.001]. Critically, the rVBI varied depending on the credibility of feedback (F(2,609)=14.83, p<0.001), such that the rVBI for the 3-star agent was lower than that for both the 1-star (b=-0.22, t(609)=-4.41, p<0.001) and 2-start agent (b=-0.24, F(1,609)=24.74, p<0.001). Feedback with 50% and 75% credibility yielded similar rVBI values (b=0.028, t(609)=0.56,p=0.57). Finally, a positivity bias could not stem from a Bayesian strategy as both Bayesian models predicted a negativity bias (**Fig. 5b-c**; Fig. S8; and SI 3.1.1.3 Table S11-S12, 3.2.1.1, and 3.2.1.2).

Previous research has suggested that positivity bias may spuriously arise from pure choiceperseveration (i.e., a tendency to repeat previous choices regardless of outcome) (49,50). While our models included a perseveration-component, this control may not be preferent. Therefore, in additional control analyses, we generated synthetic datasets using models including choiceperseveration but devoid of feedback-valence bias, and fitted them with our credibility-valence model (see SI 3.6.1). These analyses confirmed that perseveration can masquerade as an apparent positivity bias. Critically, however, these analyses also confirmed that perseveration cannot account for our main finding of increased positivity bias, relative to the overall extent of CA, for low-credibility feedback.

True feedback elicits greater learning

Our findings are consistent with participant modulation of the extent of credit-assignment based *solely* on cued task-variables, such as feedback-credibility and valence. However, we also considered another possibility: that participants might infer, on a *trial-by-trial* basis, whether the feedback they received was true or false and adjust their credit assignment based on this inference. For example, for a given feedback-agent, participants might boost the credit assigned to a chosen bandit as a function of the degree to which they believe feedback was true. Notably, Bayesian inference can support a trial-level calculation of a posterior probability that feedback is true based on its credibility, valence and a prior belief (based on experiences in previous trials) regarding the probability that the chosen bandit is truly rewarding (**Fig. 6a**). The beliefs can partly discriminate between truthful and false feedback. These beliefs can partially discriminate between truthful and false feedback. As proof of this, we calculated a Bayesian posterior feedback-truthfulness belief for each participant and trial featuring the 1- or 2-star agents, (Methods; Recall for the 3-star agent, feedback is always true). On testing whether these posterior-truthfulness beliefs vary as a function of objective feedback truthfulness (true vs. lie), we found beliefs are stronger for truthful trials than for untruthful trials for both agents (1-star agent: mean difference=0.10, t(203)=39.47, p<0.001; 2-star agent: mean difference=0.08 , t(203)=34.43, p<0.001) (**Fig. 6b** and Fig. S9a). Note that this calculation was feasible because, as experimenters, we had privileged access to the objective truth of the choice-feedback as, when designing the experimental sessions, we generated latent true choice outcomes which could be compared to agent-reported feedback.

To formally address whether feedback truthfulness modulates credit assignment, we fitted a new variant of the CA model (the "Truth-CA" model) to the data. This variant works as our Credibility-CA model, but incorporated a truth-bonus parameter (*TB*) which increases the degree of credit assignment for feedback as a function of the *experimenter-determined* likelihood the feedback is

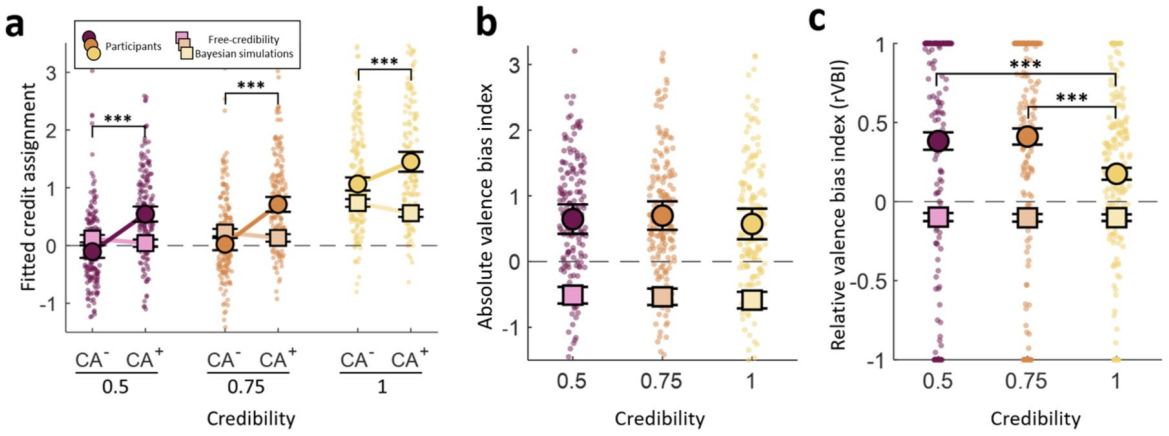

Figure 5

Positivity bias as a function of agent-credibility.

a, ML parameters from the credibility-valence-CA model. CA+ and CA- are free parameters representing credit assignments for positive and negative feedback respectively (for each credibility level). Our data revealed a positivity bias (CA+ > CA-) for all credibility levels. **b,** Absolute valence bias index (defined as CA^{+}-CA^{-}) based on the ML parameters from the credibility-valence CA model. Positive values indicate a positivity bias, while negative values represent a negativity bias. **c,** Relative valence bias index (defined as $(CA^{+}\text{-}CA^{-})/(|CA^{+}|+|CA^{-}|)$) based on the ML parameters from the credibility-valence CA model. Positive values indicate a positivity bias, while negative values represent a negativity bias. Small dots represent fitted parameters for individual participants and big circles represent the group median (a,b) or mean (c) (both of participants' behavior), while squares are the median or mean of the fitted parameters of the free-credibility Bayesian model simulations. Error bars show the standard error of the mean. (***) p<.001 for ML fits of participants behavior.

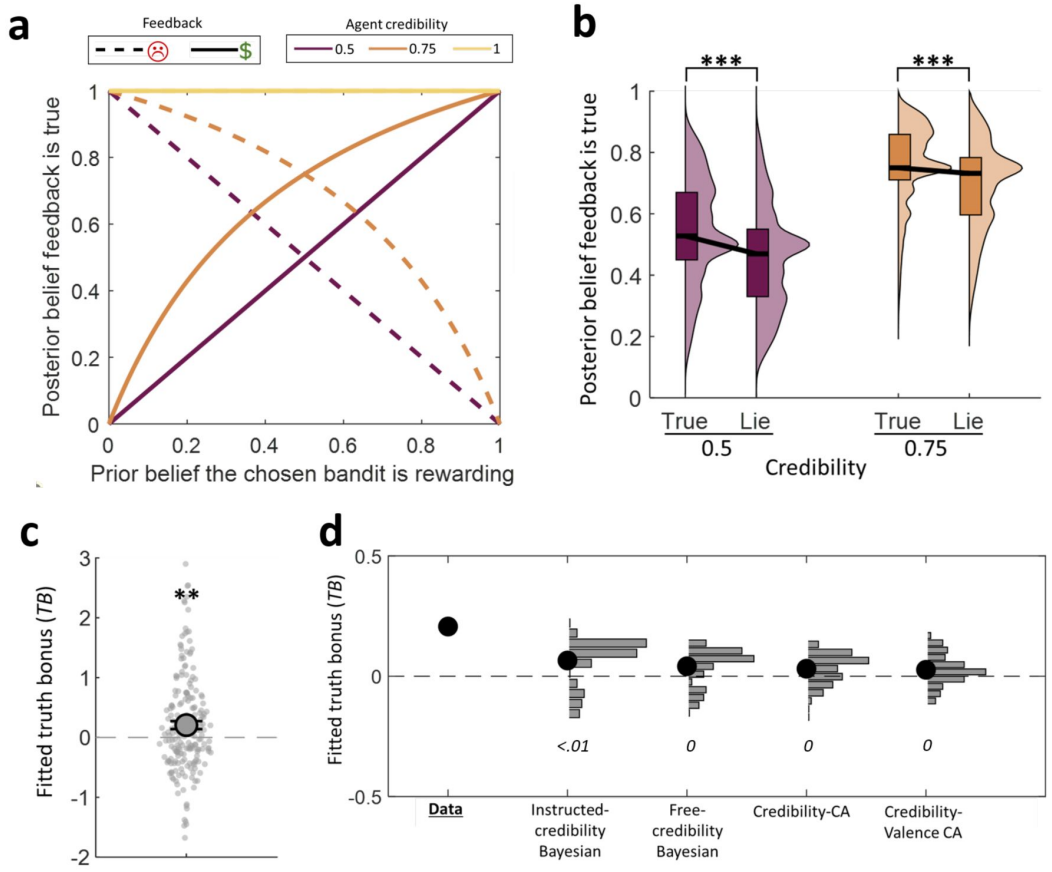

Figure 6

Credit assignment is enhanced for feedback that is more likely to be true.

a, The posterior belief that the received feedback is truthful (y-axis) is plotted against the prior belief (held before receiving feedback) that the chosen bandit would be rewarding (x-axis). The plot illustrates how this posterior belief is influenced by the valence of the feedback (reward indicated by solid lines, no reward by dashed lines) and the credibility of the feedback agent (represented by different colors). **b,** Distribution of posterior belief probability that feedback is true, calculated separately for each agent (1 or 2 star) and objective feedback-truthfulness (true or lie). These probabilities were computed based on trial-sequences and feedback participants experienced, indicating belief probabilities that feedback is true are higher in truth compared to lie trials. For illustration, plotted distributions pool trials across participants. The black line within each box represents the median, upper and lower bounds represent the third and first quartile respectively. The width of each half-violin plot corresponds to the density of each posterior belief value among all trials for a given condition. **c,** Maximum likelihood (ML) estimate of the "truth-bonus" parameter derived from the "Truth-CA" model. The significantly positive truth bonus indicates that participants increased credit assignment as a function of the likelihood this feedback was true (after controlling for the credibility of this feedback) . Each small dot represents the fitted truth-bonus parameter for an individual participant, the large circle indicates the group mean, and the error bars represent the standard error of the mean. **d**, Distribution of truth-bonus parameters predicted by synthetic simulations of our alternative computational models. For each alternative model, we generated 101 synthetic group-level datasets based on the maximum likelihood parameters fitted to the participants' actual behavior. Each of these datasets was then independently fitted with the "Truth-CA" model. Each histogram represents the distribution of the mean truth bonus across the 101 simulated group-level datasets for a specific alternative model. Notably, the truth bonus observed in our participants was significantly higher than the truth bonus predicted by any of these alternative models (proportion of datasets predicting a higher truth bonus: Instructed-credibility Bayesian < 0.01, Free-credibility Bayesian = 0, Credibility-CA = 0, Credibility-Valence CA = 0). (**) p<.01

true (which is read from the curves in Fig 6a when x is taken to be the true probability the bandit is rewarding). Specifically, after receiving feedback, the Q-value of the chosen option is updated according to the following rule:

$$Q \leftarrow \left(1 - f_Q\right) * Q + [CA(agent) + TB * (P(truth) - 0.5)] * F$$

where *TB* is the free parameter representing the truth bonus, and *P(truth)* is the probability the received feedback being true (from the experimenter's perspective). We acknowledge that this model falls short of providing a mechanistically plausible description of the credit assignment process, because, participants have no access to the experimenter's truthfulness likelihoods (as the true bandit reward probabilities are unknown to them). Nonetheless, we use this 'oracle model' as a measurement tool to glean rough estimates for the extent to which credit assignment Is boosted as a function of its truthfulness likelihood.

Fitting this Truth-CA model to participants' behaviour revealed a significant positive truth-bonus (mean=0.21, t(203)=3.12, p=0.002), suggesting that participants indeed assign greater weight to feedback that is likely to be true (**Fig. 6c**; see SI 3.3.1 for detailed ML parameter results). Notably, simulations using our other models (Methods) consistently predicted smaller truth biases (compared to the empirical bias) (**Fig. 6d**). Moreover, truth bias was still detected even in a more flexible model that allowed for both a positivity bias and truth-bias (see SI 3.7). The upshot is that participants are biased to assign higher credit based on feedback that is more likely to be true in a manner that is inconsistent with out Bayesian models and above and beyond the previously identified positivity biases.

Discovery study

The discovery study (n=104) used a disinformation task structurally similar to that used in our main study, but with three notable differences: 1) it included 4 feedback agents, with credibilities of 50%, 70%, 85% and 100%, represented by 1, 2, 3, and 4 stars, respectively; 2) each experimental block consisted of a single bandit pair, presented over 16 trials (with 4 trials for each feedback agent); and 3) in certain blocks, unbeknownst to participants, the two bandits within a pair were equally rewarding (see SI section 1.1). Overall, this study's results supported similar conclusions as our main study (see SI section 1.2) with a few differences. We found convergent support for increased learning from more credible sources (SI 1.2.1), superior fit for the CA model over Bayesian models (SI 1.2.2) and increased learning from feedback inferred to be true (SI 1.2.6). Additionally, we found an inflation of positivity bias for low-credibility both when measured relative to the overall level of credit assignment (as in our main study), or in absolute terms (unlike in our main study) (Fig. S3; SI 1.2.5). Moreover, choiceperseveration could not predict an amplification of positivity bias for low-credibility sources (see SI 3.6.2). However, we found no evidence for learning based on 50%-credibility feedback when examining either the feedback effect on choice repetition or CA in the credibility-CA model (SI 1.2.3).

Discussion

Accurate information enables individuals to adapt effectively to their environment (51,52). Indeed, it has been suggested that the importance and utility of information elevate its status to that of a secondary reinforcer, imbuing it with intrinsic value beyond its immediate usefulness (53,54). However, a significant societal challenge arises from the fact that, as social animals, much information we receive is mediated by others, entailing it can be inaccurate, biased or purposefully misleading. Here, using a novel variant of the two-armed bandit task, we asked how we update our beliefs in the presence of potential disinformation, wherein *true choice* outcomes are latent and feedback is provided by potentially disinformative agents.

We acknowledge that several factors may limit the external validity of our task, including the fact that participants were explicitly instructed about the credibility of information sources. In contrast, in many real-life scenarios, individuals need to learn the credibility of information sources based on their own experience of the world or may even have false beliefs regarding the source-credibility of agents. Moreover, in our task, the experimenter fully controlled the credibility of the information source in every trial, whereas in many real-life situations people can exercise a degree of control over the credibility of information they receive. For example, search engines allow an exercise of choice regarding the credibility of sources. Finally, in our task, feedback agents served as rudimentary representations of social agents, who lied randomly and arbitrarily, in a motivation-free manner. Conversely, in real life, others may strategically attempt to mislead us, and we can exploit knowledge of their motivation to lie, such as when we assume that a used cars seller is more likely to portray a clapped-out car as excellent, rather than state the unfiltered truth. Nevertheless, our results attest to the utility of our task in identifying biased aspects of learning in the face of disinformation, even in a simplified scenario.

Consistent with Bayesian-learning principles, we show that individuals increased their learning as a function of feedback credibility. This aligns with previous studies demonstrating an impressive human ability to flexibly increase learning rates when environmental changes render prior knowledge obsolete (23,55,56), and when there is reduced inherent uncertainty, such as "observation noise" (23,55–57). However, as hypothesized, when facing potential disinformation, we also find that individuals exhibit several important biases i.e., deviations from strictly idealized Bayesian strategies. Future studies should explore if and under what assumptions, about the task's generative structure and/or learner's priors and objectives, more complex Bayesian models (e.g., active inference (58)) might account for our empirical findings. In our main study, we show that participants revised their beliefs based on entirely non-credible feedback, whereas an ideal Bayesian strategy dictates such feedback should be ignored. This finding resonates with the "continued-influence effect" whereby misleading information continues to influence an individual's beliefs even after it has been retracted (59,60). One possible explanation is that some participants failed to infer that feedback from the 1- star agent was statistically void of information content, essentially random (e.g., the group-level credibility of this agent was estimated by our free-credibility Bayesian model as higher than 50%). Participants were instructed that this feedback would be "a lie" 50% of the time but were not explicitly told that this meant it was random and should therefore be disregarded. Notably, however, there was no corresponding evidence random feedback affected behaviour in our discovery study. It is possible that an individual's ability to filter out random information might have been limited due to a high cognitive load induced by our main study task, which required participants to track the values of three bandit pairs and juggle between three interleaved feedback agents (whereas in our discovery study each experimental block featured a single bandit pair). Future studies should explore more systematically how the ability to filter random feedback depends on cognitive load (61).

Previous reinforcement learning studies, report greater credit-assignment based on positive compared to negative feedback, albeit only in the context of veridical feedback (43,44,62). Here, supporting our a-priori hypothesis we show that this positivity bias is amplified for information of low and intermediate credibility (in absolute terms in the discovery study, and relative to the overall extent of CA in both studies) . Of note, previous literature has interpreted enhanced learning for positive outcomes in reinforcement learning as indicative of a confirmation bias (42,44). For example, positive feedback may confirm, to a greater extent than negative feedback one's choice as superior (e.g., "I chose the better of the two options"). Leveraging the framework of motivated cognition (35), we posited that feedback of uncertain veracity (e.g., low credibility) amplifies this bias by incentivising individuals to self-servingly accept positive feedback as true (because it confers positive, desirable outcomes), and explain away undesirable, choice-disconfirming, negative feedback as false. This could imply an amplified confirmation bias on social media, where content from sources of uncertain credibility, such as unknown or

unverified users, is more easily interpreted in a self-serving manner, disproportionately reinforcing existing beliefs (63). In turn, this could contribute to an exacerbation of the negative social outcomes previously linked to confirmation bias such as polarization (64,65), the formation of 'echo chambers' (19), and the persistence of misbelief regarding contemporary issues of importance such as vaccination (66,67) and climate change (68–71). We note however, that further studies are required to determine whether positivity bias in our task is indeed a form of confirmation bias. A striking finding in our study was that for a fully credible feedback agent, credit assignment was exaggerated (i.e., higher than predicted by our Bayesian models). Furthermore, the effect of fully credible feedback on choice was further boosted when it was preceded by a low-credibility context related to current learning. We interpret this in terms of a "contrast effect", whereby veridical information looms larger against a backdrop of disinformation (21). One upshot is that exaggerated learning might entail a risk of jumping to premature conclusions based on limited credible evidence (e.g., a strong conclusion that a vaccine produces significant side-effect risks based on weak credible information, following non-credible information about the same vaccine). An intriguing possibility, that could be tested in future studies, is that participants strategically amplify the extent of learning from credible feedback to dilute the impact of learning from non-credible feedback. For example, a person scrolling through a social media feed, encountering copious amounts of disinformation, might amplify the weight they assign to credible feedback in order to dilute effects of 'fake news'. Ironically, these results also suggest that public campaigns might be more effective when embedding their messages in low-credibility contexts, which may boost their impact.

Our findings show that individuals increase their credit assignment for feedback in proportion to the perceived probability that the feedback is true, even after controlling for source credibility and feedback valence. Strikingly, this learning bias was not predicted by any of our Bayesian or creditassignment (CA) models. Notably, our evidence for this bias is based on a "oracle model" that incorporates the probability of feedback truthfulness from the experimenter's perspective, rather than the participant's. This raises an important open question: how do individuals form beliefs about feedback truthfulness, and how do these beliefs influence credit assignment? Future research should address this by eliciting trial-by-trial beliefs about feedback truthfulness. Doing so would also allow for testing the intriguing possibility that an exaggerated positivity bias for non-credible sources reflects, to some extent, a truth-based discounting of negative feedback—i.e., participants may judge such feedback as less likely to be true. However, it is important to note that the positivity bias observed for fully credible sources (here and in other literature) cannot be attributed to a truth bias—unless participants were, against instructions, distrustful of that source.

An important question arises as to the psychological locus of the biases we uncovered. Because we were interested in how individuals process disinformation—deliberately false or misleading information intended to deceive or manipulate—we framed the feedback agents in our study as deceptive, who would occasionally "lie" about the true choice outcome. However, statistically (though not necessarily psychologically), these agents are equivalent to agents who mix truth-telling with random "guessing" or "noise" where inaccuracies may arise from factors such as occasionally lacking access to true outcomes, simple laziness, or mistakes, rather than an intent to deceive. This raises the question of whether the biases we observed are driven by the perception of potential disinformation as deceitful per se or simply as deviating from the truth. Future studies could address this question by directly comparing learning from statistically equivalent sources framed as either lying or noisy. Unlike previous studies wherein participants had to infer source credibility from experience (30,37,72), we took an explicit-instruction approach, allowing us to precisely assess source-credibility impact on learning, without confounding it with errors in learning about the sources themselves. More broadly, our work connects with prior research on observational learning, which examined how individuals learn from the actions or advice of social partners (72–75). This body of work has demonstrated that individuals integrate learning from their private experiences with learning based on others' actions or advice—whether by inferring the value others attribute to different options or by mimicking their behavior

(57,76). However, our task differs significantly from traditional observational learning. Firstly, our feedback agents interpret outcomes rather than demonstrating or recommending actions (30,37,72). Secondly, participants in our study lack private experiences unmediated by feedback sources. Finally, unlike most observational learning paradigms, we systematically address scenarios with deliberately misleading social partners. Future studies could bridge this by incorporating deceptive social partners into observational learning, offering a chance to develop unified models of how individuals integrate social information when credibility is paramount for decision-making.

We conclude by noting previous research has often attributed the negative impacts of disinformation, such as polarization and the formation of echo chambers, to intricate processes facilitated by external or self-selection of information (77–79). These processes include algorithms tailoring information to align with users' attitudes (80) or individuals consciously opting to engage with like-minded peers (81). However, our study reveals a more profound effect of disinformation, namely that even in minimal conditions, when low credibility information is explicitly identified, disinformation significantly impacts individuals' beliefs and decision-making processes. This occurs even when the decision at hand entails minimal emotional engagement or pertinence to deep, identity-related, issues. A critical next step is to deepen our understanding of these biases, particularly within complex social environments, not least to enable the development of effective prospective interventions capable of mitigating the potentially pernicious impacts of disinformation.

Materials and methods

Participants

We recruited 246 participants (mean age 39.33± 12.65, 112 female) from the Prolific participant pool (*www.prolific.co*) who went on to perform the task on the Gorilla platform (82). All participants were fluent English speakers with normal or corrected-to-normal vision and a Prolific approval rate of 95% or higher. UCL Research Ethics Committee approved the study (Project ID 6649/004), and all participants provided prior informed consent.

Experimental protocol

Traditional two-armed bandit task

At the beginning of the experiment participants completed a traditional version of the two-armed bandit task. Participants performed 45 trials, each featuring one of three randomly interleaved bandit pairs (such that each pair was presented on 15 trials). On each trial, participants choose between the bandit-pair, with each bandit being represented by a distinct identicon. Once a bandit was selected it generated a true outcome (converted to bonus monetary compensation) corresponding to either a reward or nothing. Within each bandit-pair, one bandit provided rewards on 75% of trials (with 25% providing no-reward), while the other bandit rewarded on 25% of the trials (75% non-reward trials). Participants were uninformed about the reward probabilities of each bandit and had to learn these based on experience.

At onset of each trial, the two bandits were presented, one on each side of the screen, and participants were asked to indicate their choice within 3 seconds by pressing the left/right arrow-keys. If the 3 seconds elapsed with no choice, participants were shown a "too slow" message and proceeded to the next trial. Following choice, the unselected bandit disappeared, and the participants were presented with the outcome of the selected bandit for 1200ms, followed by a 250

ms ISI before the start of the next trial. Rewards were represented by a green dollar symbol and non-rewards by a red sad face (both in the center of the screen). At the end of the task, participants were informed about the number of rewards they had earned.

Disinformation task

This involved a modified, disinformation version, of the same two-armed bandit task. Participants performed 8 blocks, each consisting of 45 trials. Each block followed the structure of the traditional two-armed bandit task, but with a critical difference: true choice-outcomes were withheld from participants and instead they received reward-feedback from a feedback agent. Participants were instructed prior to the task that feedback agents mostly provide accurate feedback (i.e., the true outcome) but could lie on a random minority of trials by reporting a reward in case of a true nonreward, or vice versa. The task featured three feedback agents varying in their credibility (i.e., probability of truth-telling), as indicated by a "star-rating" system, about which participants were instructed prior to the task. The 3-star agent always told the truth, whereas the other 2 agents were partially credible, reporting the truth on 75% (2-star) or 50% (1-star) of the trials. Feedback agents were randomly interleaved across trials subject to the constraint that each agent appeared on 5-trials for each bandit pair.

At the onset of each trial, participants were presented with the feedback agent for the trial (screen center) and with the two bandits, one on each side of the screen. Participants made a 2-second time limited choice by pressing the left/right arrow-keys. Following choice, the unselected bandit disappeared, and were then presented with the agent feedback for 1200ms (represented by either a rewarding green dollar sign or a non-rewarding red sad face in the center of the screen). All stimuli then disappeared for 250 ms to be followed by the start of the next trial. At the end of each block, participants were informed about the number of true rewards they had earned. They then received a 30-second break before the next block started with new 3 bandit pairs.

General protocol

At the beginning of the experiment, participants were presented with instructions for the traditional two-armed bandit task. The instructions were interleaved with four multiple-choice questions. When participants answered a question incorrectly, they could re-read the instructions and re-attempt. If participants answered a question incorrectly twice, they were compensated for the time but could not continue to the next stage. Upon completing the instructions participants proceeded to the traditional two-armed bandit task.

After the two-armed bandit task, participants were presented with instructions regarding the disinformation task. Again, these were interleaved with six questions wherein participants had two attempts to answer each question correctly. If they answered a question incorrectly twice, they were rejected and received partial participatory compensation. Participants then proceeded to the disinformation task. After completing the disinformation task, participants completed three psychiatric questionnaires (presented in random order): 1) the Obsessional Compulsive Inventory - Revised (OCI-R) (83), assessing symptoms of obsessive-compulsive disorder (OCD); 2) The Revised Green et al. Paranoid Thoughts Scale (R-GPTS) (84), measuring paranoid ideations; and 3) the DOG scale, evaluating dogmatism (85).

The participants took on average 43 minutes to complete the experiment. They received a fixed compensation of 5.48 GBP and variable compensation between 0 and 2 GBP based on their performance on the disinformation task.

Attention checks

The two tasks included randomly interleaved catch trials wherein participants were cued to press a given key within a 3-second limit. None of the participants failed more than one of these attention checks.

Data analysis

Exclusion criteria

Participants were excluded if they: 1) Either repeated or alternated key presses in more than 70% of the trials, and/or 2) their reaction time was lower than 150 ms in more than 5% of the trials. Based on these criteria 42 participants were excluded, while 204 participants were kept for the analyses.

Accuracy

Accuracy rates were calculated as the probability of choosing within a given pair the bandit with a higher reward probability. For **figure 1d**, we calculated for each participant and for each trial (within a bandit-pair) averaged accuracy across all bandit-pairs. We then averaged accuracy at the trial level across participants. Overall improvement for each participant was calculated as the average accuracy difference between the last and first trials for each of the bandit-pairs.

Computational models

RL Models

We formulated a family of RL models to account for participant choices. In these models, a tendency to choose each bandit is captured by a Q-value. After reward-feedback the Q-value of the chosen bandit was updated conditional on the agent and on whether the feedback was positive or negative according to the following rule:

$$Q(chosen) \leftarrow \left(1 - f_Q\right) * Q(chosen) + CA(agent, valence) * F \quad (1)$$

where *CA* is a free credit assignment parameter representing the magnitude of the value increase/decrease following feedback receipt *F* from the agents (coded as 1 for reward feedback and −1 for non-reward feedback), while *fQ* ($\in$[0,1]) is the free parameter representing the forgetting rate of the Q-value. Additionally, the value of each of the other bandits (i.e., the unchosen bandit in the presented pair and all the bandits from the other not-shown pairs) were forgotten as per the following:

$$Q(non - chosen) \leftarrow \left(1 - f_Q\right) * Q(non - chosen) \quad (2)$$

Alternative model-variants differed based on whether the CA parameter(s) were influenced by agents and/or feedback valence (see **Table 1** below), allowing us to test how these variables impacted learning.

1. The “Null” model included a unique CA parameter conveying an assumption that feedback is modulated by neither agent-credibility nor feedback valence.

2. The “Credibility-CA” models included a dedicated CA parameter for each agent allowing for the possibility learning was selectively modulated by agent credibility (but not by feedback valence).
3. The “Credibility-Valence-CA” model included distinct CA parameters for rewarding (CA+) and nonrewarding feedback (CA-) for each agent, allowing CA to be influenced by both feedback valence and credibility.
4. The “constant feedback-valence bias” CA model included separate CA-parameters for each agent, but a single valence bias parameter (VB) common to all agents, such that the CA+ parameter for each agent corresponded to the sum of its CA-parameter and the common VB parameter.

Additionally, we formulated a “Truth-CA” model, which worked as our Credibility-CA model, but incorporated a free truth-bonus parameter (*TB*). This parameter modulates the extent of credit assignment for each agent based on the posterior probability of feedback being true (given the credibility of the feedback agent, and the true reward probability of the chosen bandit). The chosen bandit was updated as follows:

$$Q \leftarrow \left(1 - f_Q\right) * Q + [CA(agent) + TB * (P(truth) - 0.5)] * F \quad (3)$$

where P(truth) is the posterior probability of the feedback being true in the current trial (for exact calculation of P(truth) see “Methods: Bayesian estimation of posterior belief that feedback is true”).

All models also included gradual perseveration for each bandit. In each trial the perseveration values (P) were updated according to

$$P(chosen) \leftarrow (1 - f_P) * P(chosen) + PERS \quad (4)$$

Where PERS is a free parameter representing the P-value change for the chosen bandit, and f_p ($\in$ [0,1]) is the free parameter denoting the forgetting rate applied to the P value. Additionally, the P-values of all the non-chosen bandits (i.e., again, the unchosen bandit of the current pair, and all the bandits from the not-shown pairs) were forgotten as follows:

$$P(non - chosen) \leftarrow (1 - f_P) * P(non - chosen) \quad (5)$$

We modelled choices using a *softmax* decision rule, representing the probability of the participant to choose a given bandit over the alternative:

$$P(bandit) = \frac{1}{1 + e^{[Q(other\ bandit) - Q(bandit)] + [P(other\ bandit) - P(bandit)]}} \quad (6)$$

Model	Free CA parameter
Null	CA
Credibility-CA	$CA_{0.5}, CA_{0.75}, CA_1$
Credibility-Valence-CA	$CA_{0.5}+, CA_{0.75}+, CA_1 +$ $CA_{0.5}-, CA_{0.75}-, CA_1 -$
constant feedback-valence bias CA	VB $CA_{0.5}-, CA_{0.75}-, CA_1 -$
Truth-CA	TB $CA_{0.5}, CA_{0.75}, CA_1$

Table 1

summary of free parameters for each of the CA models.

Bayesian Models

We also formulated a Bayesian model corresponding to an ideal belief updating strategy. In this model, beliefs about each bandit were represented by a density distribution over the probability that a bandit provides a true reward *g(p)*, where *p* is the probability of a true reward (see full derivation in SI 4.1). During learning, following reward-feedback, the distribution *for* the chosen bandit was updated based on the agent's feedback (*F*) and its associated credibility (*C*):

$$g(p) \leftarrow g(p) * [C * p + (1 - C) * (1 - p)] \quad if\ F = 1 \tag{7}$$

$$g(p) \leftarrow g(p) * [(1 - C) * p + C * (1 - p)] \quad if\ F = -1 \tag{8}$$

$$g(p) \leftarrow \frac{g(p)}{\int_0^1 g(p)\, dp} \tag{9}$$

At the beginning of each block priors for each bandit were initialized to uniform distributions (*g(p)=U[0,1]*). In the *instructed-credibility Bayesian model*, we fixed the credibilities to their true values (i.e., 0.5, 0.75 and 1).

We also formulated a *free-credibility Bayesian model*, where we only fixed the three-star agent credibility to 1 but estimated the credibility of the two lying agents as free parameters. This model allowed the possibility that participants use distorted instructed-credibilities when following a Bayesian strategy.

For both versions, we modelled choice using a SoftMax function with a free inverse temperature parameter (β):

$$P(bandit) = \frac{1}{1 + e^{\beta * [Q(other\ bandit) - Q(bandit)]}} \tag{10}$$

Where here *Q(bandit)* is the expected probability, the bandit provides a true reward.

Additionally, we formulated extended Bayesian models to account for choice-perseveration (see SI 3.6.1). These models operate as our instructed-credibility and free-credibility Bayesian models, but also incorporate a perseveration values, updated in each trial as in our CA models (Eqs. **3** and **5**). For these extended models, we modelled choices using the following *softmax* decision rule:

$$P(bandit) = \frac{1}{1 + e^{\beta * [Q(other\ bandit) - Q(bandit)] + [P(other\ bandit) - P(bandit)]}} \tag{11}$$

Parameter optimization, model selection and synthetic model simulations

For each participant, we estimated the free parameter values that maximized the summed loglikelihood of the observed choices across all games. Trials where participants showed a response time below 150 ms were excluded from the log-likelihood calculations. To minimise the

chances of finding local minima, we ran the fitting procedure 10 times for each participant, using random initializations for the parameters (CA~U[-10,10], PERS~U[-5,5], f_Q~[0,1], fp~[0,1], TB~[-10,10], β~[O,3O], C~U[0,1]).

We performed model comparison between Bayesian and CA models using the parametric bootstrap cross-fitting method (PBCM)(86,87). In brief, this method relies on generating, for each participant, synthetic datasets (we used 201) based on maximal likelihood parameters and each model variant (i.e., the Bayesian model and the CA model), and fitting each dataset with the two models. We then calculated the log likelihood difference between the two fits for each dataset, obtaining two loglikelihood difference distributions, one for each generative model. We determined a loglikelihood difference threshold that leads to best model-classification (i.e., maximizing the proportion of true positives and true negatives). Finally, we fit the empirical data from each participant with the two model variants, calculating an empirical loglikelihood difference. A comparison of this empirical likelihood difference to the classification threshold determines which model provides a better fit for a participant's data (see Fig. S6 for more information). We used this procedure to compare our Bayesian models (instructed-credibility and free-credibility Bayesian) with a simplified version of the credibility-CA model that did not include perseveration (PERS, fP = 0).

We also performed model-comparisons for nested CA models using generalized-likelihood ratio tests where the null distribution for rejecting a nested model (in favour of a nesting model) was based on a bootstrapping method (BGLRT)(48,88).

To assess the mechanistic predictions of each model, we generated synthetic simulations based on the ML parameters of participants. Unless stated otherwise, we generated 5 simulations for each participant (1020 total simulations) with a new sequence of trials generated as in the actual data. We analysed these data in the same way as we analysed empirical data, after pooling together the 5 simulated data set per participant.

Parameter recovery

For each model of interest, we generated 201 synthetic simulations based on parameters sampled from uniform distributions (CA~U[-10,10], PERS~U[-5,5], f_Q~U[0,1], f_p~U[0,1], β~U[0,30], C~U[0,1]). We fitted each simulated dataset with its generative model and calculated the Spearman's correlation between the generative and fitted parameters.

Mixed effects models

Model-agnostic analysis of agent-credibility effects on choice-repetition

We used a mixed-effects binomial regression model to assess whether, and how, value-learning was modulated by agent-credibility, with participants serving as random effects. The regressed variable *REPEAT* indicated whether the current trial repeated the choice from the previous trial featuring the same bandit-pair (repeated choice=1, non-repeated choice=0) and was regressed on the following regressors: *FEEDBACK* coded whether feedback received in the previous trial with the same bandit pair was positive or negative (coded as 0.5, −0.5, respectively), *BETTER* coded whether the bandit chosen in that previous trial was the better -mostly rewarding- or the worse -mostly unrewarding-bandit within the pair, coded as 0.5 and −0.5 respectively, $AGENT_{2\text{-star}}$ indicated whether feedback received in the previous trial (featuring the same bandit pair) came

from the 2-star agent (previous feedback from 2-star agent=1, otherwise=0) and, $AGENT_{3\text{-star}}$ indicated whether the feedback in the previous trial came from the 3-star agent. The model in Wilkinson's notation was:

$$REPEAT \sim FEEDBACK * BETTER * (AGENT_{2-star} + AGENT_{3-star}) + (1|participant) \quad (12)$$

In **figure 2a and 2b**, we plot the choice-repeat probability based on feedback-valence and agentcredibility from the preceding trial with the same bandit pair. We independently calculated the repeat probability for the better (mostly rewarding) and worse (mostly non-rewarding) bandits and averaged across them. This calculation was done at the participants level, and finally averaged across participants.

Model-agnostic analysis of contextual credibility effects on choice-repetition

We used a different mixed-effects binomial regression model to test whether value learning from the 3-star agent was modulated by contextual credibility. We focused this analysis on instances where the previous trial with the same bandit pair featured the 3-star agent. We regressed the variable *REPEAT*, which indicated whether the current trial repeated the choice from the previous trial featuring the same bandit-pair (repeated choice=1, non-repeated choice=0). We included the following regressors: *FEEDBACK* coding the valence of feedback in the previous trial with the same bandit pair (positive=0.5, negative=-0.5), $CONTEXT_{2\text{-star}}$ indicating whether the trial immediately preceding the previous trial with the same bandit pair (context trial) featured the 2-star agent (feedback from 2-star agent=1, otherwise=0), and $CONTEXT_{3\text{-star}}$ indicating whether the trial immediately preceding the previous trial with the same bandit pair featured the 3-star agent. We also included a regressor (*BETTER*) coding whether the bandit chosen in the learning trial was the better -mostly rewarding- or the worse -mostly unrewarding- bandit within the pair. We included in this analysis only current trials where the context trial featured the same bandit pair. The model in Wilkinson's notation was:

$$REPEAT \sim FEEDBACK * (CONTEXT_{2-star} + CONTEXT_{3-star}) + BETTER + (1|participant) \quad (13)$$

In **figure 4c**, we independently calculate the repeat probability difference for the better (mostly rewarding) and worse (mostly non-rewarding) bandits and averaged across them. This calculation was done at the participants level, and finally averaged across participants.

Effects of agent-credibility on CA parameters from credibility-CA model

We used a mixed-effects linear regression model to assess whether, and how, credit assignment was modulated by feedback-agent, with participants serving as random effects (data from **Fig. 2c**). We regressed the maximal likelihood CA parameters from the credibility-CA model. The regressors $AGENT_{2\text{-star}}$ and $AGENT_{3\text{-star}}$ indicated, respectively, whether the CA parameter was attributed to the 2- star or the 3-star agent. The model's Wilkinson's notation was:

$$CA \sim AGENT_{2-star} + AGENT_{3-star} + (1|participant) \quad (14)$$

Effects of agent-credibility and feedback valence on CA parameters from credibility valence-CA model

We used a second mixed-effects linear regression model to test for a valence bias in learning, and how such bias was modulated by feedback credibility, with participants serving again as random effects (data from **Fig. 3a**). The maximal likelihood CA parameters from the credibility-valence-CA model served as the regressed variable, which was regressed on: $AGENT_{2\text{-star}}$ and $AGENT_{3\text{-star}}$ (defined in the same way as the previous model), and *VALENCE* coding whether the CA parameter was attributed to positive (coded as 0.5) or negative (coded as −0.5) feedback. The Wilkinson's notation of the model was:

$$CA \sim VALENCE * (AGENT_{2-star} + AGENT_{3-star}) + (1|participant) \quad (15)$$

We used a separate mixed-effects linear regression model to test how relative valence bias was modulated by feedback credibility. We first computed the relative valence bias index (rVBI) for each credibility level, and we then regressed these values on $AGENT_{2\text{-star}}$ and $AGENT_{3\text{-star}}$ (defined in the same way as the previous models).

$$rVBI = \frac{CA^+ - CA^-}{|CA^+| + |CA^-|} \quad (16)$$

$$rVBI \sim AGENT_{2-star} + AGENT_{3-star} + (1|participant)$$

Bayesian estimation of posterior belief that feedback is true

We calculated the Bayesian posterior conditional probability of feedback truthfulness (**Fig. 4a and 4b**) follows. First, we calculated the probability of each true outcome, r (0-non-reward; 1-reward) conditional on the feedback, *f* (0: non-reward, 1: reward), the credibility of the agent reporting the feedback (C) and the history of experiences from past trials (H):

$$Prob(r|f,C,H) \propto Prob(f|r,C,H) * Prob(r|C,H) = Prob(f|r,C) * p(r|H) = \left[C * 1_{f=r} + (1-C) * 1_{f\neq r}\right] * \left[\bar{p} * 1_{r=1} + (1-\bar{p}) * 1_{r=0}\right] \quad (17)$$

Where proportionality omits terms independent of r, $\bar{p} = \int_0^1 p\,g(p|H)$ is the expected probability of the chosen bandit is rewarding (conditional on past-trial history), and *g(p|H)* is the density over the probability (the chosen bandit) is rewarded (conditional on the history of previous trials).

Next, we normalized the two terms (for *r*=0,1) to sum to 1 (to correct for the proportionality in (14)). Finally, the posterior belief in truthfulness was taken as *P(r=f|f,C,H)*.

In **Fig. 4b**, we calculated for each participant the mean posterior belief of truthfulness separately for trials where each agents told the truth or lied, and we compared these mean beliefs between the two kinds of trials using a paired t-tests (one test per agent).

Code and Data Availability

All code and data used to generate the results and figures in this paper will be made available on GitHub upon publication.

Acknowledgements

We thank Bastien Blain, Lucie Charles and Stephano Palminteri for helpful discussions. We thank Nira Liberman, Keiji Ota, Nitzan Shahar, Konstantinos Tsetsos and Tali Sharot for providing feedback on earlier versions of the manuscript. We additionally thank the members of the Max Planck UCL Centre for Computational Psychiatry and Ageing Research for insightful discussions. The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society.

J.V.P. is a pre-doctoral fellow of the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research (IMPRS COMP2PSYCH). We acknowledge funding from the Max Planck research school to J.V.P. (577749-D-CON 186534), and funding from the Max Planck Society to R.J.D. (549771-D.CON 177814). The project that gave rise to these results received the support of a fellowship from “la Caixa” Foundation (ID 100010434), with the fellowship code LCF/BQ/EU21/11890109.

J.V.P. contributed to the study design, data collection, data coding, data analyses, and writing of the manuscript. R.M. contributed to the study design, data analyses, and writing of the manuscript. R.J.D. contributed to the writing of the manuscript.

Additional files

Supplementary information

References

1. World Economic Forum **Global Risks Report 2024** https://www.weforum.org/publications/global-risks-report-2024/

2. Carrieri V, Madio L, Principe F (2019) **Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy** *Health Econ* **28**:1377–82 Google Scholar

3. Rocha YM, de Moura GA, Desidèrio GA, de Oliveira CH, Lourengo FD, de Figueiredo Nicolete LD (2023) **The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review** *J Public Health* **31**:1007–16 Google Scholar

4. Belluz J. Vox (2017) **Why Japan's HPV vaccine rates dropped from 70% to near zero** https://www.vox.com/science-and-health/2017/12/1/16723912/japan-hpv-vaccine

5. Horta Ribeiro M, Calais PH, Almeida VAF, Meira W (2017) **"Everything I Disagree With is #FakeNews": Correlating Political Polarization and Spread of Misinformation. arXiv eprints** https://ui.adsabs.harvard.edu/abs/2017arXiv170605924H

6. Piazza JA (2022) **Fake news: the effects of social media disinformation on domestic terrorism** *Dyn Asymmetric Confl* **15**:55–77 Google Scholar

7. Roy S, Singh AK, Kamruzzaman (2023) **Sociological perspectives of social media, rumors, and attacks on minorities: Evidence from Bangladesh** *Front Sociol* **8**:1067726 Google Scholar

8. BBC Trending (2016) **The saga of "Pizzagate": The fake story that shows how conspiracy theories spread** *BBC News* https://www.bbc.com/news/blogs-trending-38156985

9. Enders AM, Uscinski JE, Seelig MI, Klofstad CA, Wuchty S, Funchion JR, et al. (2023) **The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation** *Polit Behav* **45**:781–804 Google Scholar

10. Guess A, Nagler J, Tucker J (2019) **Less than you think: Prevalence and predictors of fake news dissemination on Facebook** *Sci Adv* **5**:eaau4586 Google Scholar

11. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. (2016) **The spreading of misinformation online** *Proc Natl Acad Sci* **113**:554–9 Google Scholar

12. Shao C, Ciampaglia GL, Varol O, Yang KC, Flammini A, Menczer F (2018) **The spread of low-credibility content by social bots** *Nat Commun* **9**:4787 Google Scholar

13. Sharevski F, Alsaadi R, Jachim P, Pieroni E (2022) **Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes** *Comput Secur* **114**:102577 Google Scholar

14. Walter N, Murphy ST (2018) **How to unring the bell: A meta-analytic approach to correction of misinformation** *Commun Monogr* **85**:423–41 Google Scholar

15. Globig LK, Holtz N, Sharot T (2022) **Changing the Incentive Structure of Social Media Platforms to Halt the Spread of Misinformation** https://psyarxiv.com/26j8w/

16. Roozenbeek J, van der Linden S (2019) **Fake news game confers psychological resistance against online misinformation** *Palgrave Commun* **5**:1–10 Google Scholar

17. O'Mahony C, Brassil M, Murphy G, Linehan C (2023) **The efficacy of interventions in reducing belief in conspiracy theories: A systematic review** *PLOS One* **18**:e0280902 Google Scholar

18. Vosoughi S, Roy D, Aral S (2018) **The spread of true and false news online** *Science* **359**:1146–51 Google Scholar

19. Modgil S, Singh RK, Gupta S, Dennehy D (2021) **A Confirmation Bias View on Social Media Induced Polarisation During Covid-19** *Inf Syst Front* https://doi.org/10.1007/s10796-021-10222-9 | Google Scholar

20. Menczer F, Ciampaglia GL (2018) **The Conversation** *Misinformation and biases infect social media, both intentionally and accidentally* http://theconversation.com/misinformation-and-biases-infect-social-media-both-intentionally-and-accidentally-97148

21. Pennycook G, Bear A, Collins ET, Rand DG (2020) **The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings** *Manag Sci* **66**:4944–57 Google Scholar

22. Swire B, Berinsky AJ, Lewandowsky S, Ecker UKH (2017) **Processing political misinformation: comprehending the Trump phenomenon** *R Soc Open Sci* **4**:160802 Google Scholar

23. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) **Learning the value of information in an uncertain world** *Nat Neurosci* **10**:1214–21 Google Scholar

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

25. Diederen KMJ, Schultz W (2015) **Scaling prediction errors to reward variability benefits error-driven learning in humans** *J Neurophysiol* **114**:1628 Google Scholar

26. Campbell-Meiklejohn D, Simonsen A, Frith CD, Daw ND (2017) **Independent Neural Computation of Value from Other People's Confidence** *J Neurosci* **37**:673–84 Google Scholar

27. De Martino B, Bobadilla-Suarez S, Nouguchi T, Sharot T, Love BC (2017) **Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability** *J Neurosci* **37**:6066–74 Google Scholar

28. Toelch U, Bach DR, Dolan RJ (2014) **The neural underpinnings of an optimal exploitation of social information under uncertainty** *Soc Cogn Affect Neurosci* **9**:1746–53 Google Scholar

29. Biele G, Rieskamp J, Gonzalez R (2009) **Computational models for the combination of advice and individual learning** *Cogn Sci* **33**:206–42 Google Scholar

30. Velez N, Gweon H (2019) **Integrating Incomplete Information With Imperfect Advice** *Top Cogn Sci* **11**:299–315 Google Scholar

31. Jiwa M, Yu Y, Boonyaratvej J, Ciston A, Haggard P, Charles L, et al. (2023) **Exposure to misleading and unreliable information reduces active information-seeking** *PsyArXiv* https://osf.io/preprints/psyarxiv/4zkxw/ | Google Scholar

32. Sharot T (2011) **The optimism bias** *Curr Biol* **21**:R941–5 Google Scholar

33. Sharot T, Garrett N (2016) **Forming Beliefs: Why Valence Matters** *Trends Cogn Sci* **20**:25–33 Google Scholar

34. Sharot T, Korn CW, Dolan RJ (2011) **How unrealistic optimism is maintained in the face of reality** *Nat Neurosci* **14**:1475–9 Google Scholar

35. Hughes BL, Zaki J (2015) **The neuroscience of motivated cognition** *Trends Cogn Sci* **19**:62–4 Google Scholar

36. Sutton RS, Barto AG (2018) **Reinforcement Learning: An Introduction** MIT Press Google Scholar

37. Schulz L, Schulz E, Bhui R, Dayan P (2023) **Mechanisms of Mistrust: A Bayesian Account of Misinformation Learning** OSF https://osf.io/8egxh

38. Aston AT (2022) **Modeling the Social Reinforcement of Misinformation Dissemination on Social Media** *J Behav Brain Sci* **12**:533–47 Google Scholar

39. Aymanns C, Foerster J, Georg CP, Weber M (2022) **Fake News in Social Networks** https://papers.ssrn.com/abstract=4173312

40. Lindstrom B, Bellander M, Schultner DT, Chang A, Tobler PN, Amodio DM (2021) **A computational reward learning account of social media engagement** *Nat Commun* **12**:1311 Google Scholar

41. Vidal-Perez J, Dolan RJ, Moran R (2025) **Biased Misinformation Distorts Beliefs** OSF https://osf.io/rk52q_v1

42. Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ (2017) **Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing** *PLOS Comput Biol* **13**:e1005684 Google Scholar

43. Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S (2017) **Behavioural and neural characterization of optimistic reinforcement learning** *Nat Hum Behav* **1**:1–9 Google Scholar

44. Palminteri S, Lebreton M (2022) **The computational roots of positivity and confirmation biases in reinforcement learning** *Trends Cogn Sci* **26**:607–21 Google Scholar

45. Brady WJ, McLoughlin K, Doan TN, Crockett MJ (2021) **How social learning amplifies moral outrage expression in online social networks** *Sci Adv* **7**:eabe5641 Google Scholar

46. Wilson RC, Collins AG (2019) **Ten simple rules for the computational modeling of behavioral data** *eLife* **8**:e49547 https://doi.org/10.7554/eLife.49547 | Google Scholar

47. Reyna VF, Brainerd CJ (2023) **Numeracy, gist, literal thinking and the value of nothing in decision making** *Nat Rev Psychol* **2**:421–39 Google Scholar

48. Moran R, Dayan P, Dolan RJ (2021) **Human subjects exploit a cognitive map for credit assignment** *Proc Natl Acad Sci* **118**:e2016884118 Google Scholar

49. Sugawara M, Katahira K (2021) **Dissociation between asymmetric value updating and perseverance in human reinforcement learning** *Sci Rep* **11**:3574 Google Scholar

50. Palminteri S (2022) **Choice-Confirmation Bias and Gradual Perseveration in Human Reinforcement Learning** *Behav Neurosci* :137 Google Scholar

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

52. Niv Y (2009) **Reinforcement learning in the brain** *J Math Psychol* **53**:139–54 Google Scholar

53. Bennett D, Bode S, Brydevall M, Warren H, Murawski C (2016) **Intrinsic Valuation of Information in Decision Making under Uncertainty** *PLOS Comput Biol* **12**:e1005020 Google Scholar

54. Bromberg-Martin ES, Monosov IE (2020) **Neural circuitry of information seeking** *Curr Opin Behav Sci* **35**:62–70 Google Scholar

55. Glaze CM, Kable JW, Gold JI (2015) **Normative evidence accumulation in unpredictable environments** *eLife* **4**:e08825 https://doi.org/10.7554/eLife.08825 | Google Scholar

56. Glaze CM, Filipowicz ALS, Kable JW, Balasubramanian V, Gold JI (2018) **A bias-variance trade-off governs individual differences in on-line learning in an unpredictable environment** *Nat Hum Behav* **2**:213–24 Google Scholar

57. Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS (2008) **Associative learning of social value** *Nature* **456**:245–9 Google Scholar

58. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G (2017) **Active Inference: A Process Theory** *Neural Comput* **29**:1–49 Google Scholar

59. Johnson HM, Seifert CM (1994) **Sources of the continued influence effect: When misinformation in memory affects later inferences** *J Exp Psychol Learn Mem Cogn* **20**:1420–36 Google Scholar

60. Walter N, Tukachinsky R (2020) **A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?** *Commun Res* **47**:155–77 Google Scholar

61. Collins AGE, Frank MJ (2012) **How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis** *Eur J Neurosci* **35**:1024–35 Google Scholar

62. Chambon V, Thero H, Vidal M, Vandendriessche H, Haggard P, Palminteri S (2020) **Information about action outcomes differentially affects learning from self-determined versus imposed choices** *Nat Hum Behav* **4**:1067–79 Google Scholar

63. Westerwick A, Sude D, Robinson M, Knobloch-Westerwick S (2020) **Peers Versus Pros: Confirmation Bias in Selective Exposure to User-Generated Versus Professional Media Messages and Its Consequences** *Mass Commun Soc* **23**:510–36 Google Scholar

64. Gallo E, Langtry A (2020) **Social Networks, Confirmation Bias and Shock Elections** Cambridge University, Faculty of Economics https://doi.org/10.17863/CAM.62312 | Google Scholar

65. Lefebvre G, Deroy O, Bahrami B (2024) **The roots of polarization in the individual reward system** *Proc R Soc B Biol Sci* **291**:20232011 Google Scholar

66. Meppelink CS, Smit EG, Fransen ML, Diviani N (2019) **"I was Right about Vaccination": Confirmation Bias and Health Literacy in Online Health Information Seeking** *J Health Commun* **24**:129–40 Google Scholar

67. Malthouse E (2023) **Confirmation bias and vaccine-related beliefs in the time of COVID-19** *J Public Health* **45**:523–8 Google Scholar

68. Huang Y, Wang W (2024) **Overcoming Confirmation Bias in Misinformation Correction: Effects of Processing Motive and Jargon on Climate Change Policy Support** *Sci Commun* :10755470241229452 Google Scholar

69. Sunstein CR, Bobadilla-Suarez S, Lazzaro SC, Sharot T (2017) **How People Update Beliefs about Climate Change: Good News and Bad News** *CORNELL LAW Rev* :102 Google Scholar

70. Zhou Y, Shen L (2022) **Confirmation Bias and the Persistence of Misinformation on Climate Change** *Commun Res* **49**:500–23 Google Scholar

71. Hart PS, Nisbet EC (2012) **Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization About Climate Mitigation Policies** *Commun Res* **39**:701–23 Google Scholar

72. Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, et al. (2014) **Inferring on the Intentions of Others by Hierarchical Bayesian Learning** *PLOS Comput Biol* **10**:e1003810 Google Scholar

73. Zhang L, Gläscher J. (2020) **A brain network supporting social influences in human decision-making** *Sci Adv* **6**:eabb4159 Google Scholar

74. Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) **Neural mechanisms of observational learning** *Proc Natl Acad Sci U S A* **107**:14431–6 Google Scholar

75. Chelarescu P (2021) **Deception in Social Learning: A Multi-Agent Reinforcement Learning Perspective** *arXiv* http://arxiv.org/abs/2106.05402 | Google Scholar

76. Charpentier CJ, Iigaya K, O'Doherty JP (2020) **A Neuro-computational Account of Arbitration between Choice Imitation and Goal Emulation during Human Observational Learning** *Neuron* **106**:687–699.e7 Google Scholar

77. Garrett RK (2009) **Echo chambers online?: Politically motivated selective exposure among Internet news users1** *J Comput-Mediat Commun* **14**:265–85 Google Scholar

78. Ross Arguedas A, Robertson C, Fletcher R, Nielsen R (2022) **Echo chambers, filter bubbles, and polarisation: a literature review** *Reuters Institute for the Study of Journalism* https://ora.ox.ac.uk/objects/uuid:6e357e97-7b16-450a-a827-a92c93729a08

79. Cardenal AS, Aguilar-Paredes C, Galais C, Perez-Montoro M (2019) **Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure** *Int J Press* **24**:465–86 Google Scholar

80. Brady WJ, Jackson JC, Lindström B, Crockett MJ (2023) **Algorithm-mediated social learning in online social networks** *Trends Cogn Sci* **27**:947–60 Google Scholar

81. Bakshy E, Messing S, Adamic LA (2015) **Exposure to ideologically diverse news and opinion on Facebook** *Science* **348**:1130–2 Google Scholar

82. Anwyl-Irvine AL, Massonnie J, Flitton A, Kirkham N, Evershed JK (2020) **Gorilla in our midst: An online behavioral experiment builder** *Behav Res Methods* **52**:388–407 Google Scholar

83. Foa EB, Huppert JD, Leiberg S, Langner R, Kichic R, Hajcak G, et al. (2002) **The Obsessive-Compulsive Inventory: Development and validation of a short version** *Psychol Assess* **14**:485–96 Google Scholar

84. Freeman D, Loe BS, Kingdon D, Startup H, Molodynski A, Rosebrock L, et al. (2021) **The revised Green et al., Paranoid Thoughts Scale (R-GPTS): psychometric properties, severity ranges, and clinical cutoffs** *Psychol Med* **51**:244–53 Google Scholar

85. Altemeyer B (2002) **Dogmatic behavior among students: testing a new measure of dogmatism** *J Soc Psychol* **142**:713–21 Google Scholar

86. Wagenmakers EJ, Ratcliff R, Gomez P, Iverson GJ (2004) **Assessing model mimicry using the parametric bootstrap** *J Math Psychol* **48**:28–50 Google Scholar

87. Moran R, Keramati M, Dayan P, Dolan RJ (2019) **Retrospective model-based inference guides modelfree credit assignment** *Nat Commun* **10**:750 Google Scholar

88. Moran R, Goshen-Gottstein Y (2015) **Old processes, new perspectives: Familiarity is correlated with (not independent of) recollection and is more (not equally) variable for targets than for lures** *Cognit Psychol* **79**:40–67 Google Scholar

Author information

Juan Vidal-Perez

Max Planck Centre for Computational Psychiatry and Ageing, University College London, London, United Kingdom, Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom

For correspondence: juanvidalpe@gmail.com

Raymond J Dolan
Max Planck Centre for Computational Psychiatry and Ageing, University College London, London, United Kingdom, Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom

Rani Moran
Max Planck Centre for Computational Psychiatry and Ageing, University College London, London, United Kingdom, Department of Psychology, School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

For correspondence: rani.moran@gmail.com

Editors

Reviewer #1 (Public review):

This is a well-designed and very interesting study examining the impact of imprecise feedback on outcomes on decision-making. I think this is an important addition to the literature and the results here, which provide a computational account of several decision-making biases, are insightful and interesting.

I do not believe I have substantive concerns related to the actual results presented; my concerns are more related to the framing of some of the work. My main concern is regarding the assertion that the results prove that non-normative and non-Bayesian learning is taking place. I agree with the authors that their results demonstrate that people will make decisions in ways that demonstrate deviations from what would be optimal for maximizing reward in their task under a strict application of Bayes rule. I also agree that they have built reinforcement learning models which do a good job of accounting for the observed behavior. However, the Bayesian models included are rather simple- per the author descriptions, applications of Bayes' rule with either fixed or learned credibility for the feedback agents. In contrast, several versions of the RL models are used, each modified to account for different possible biases. However more complex Bayes-based models exist, notably active inference but even the hierarchical gaussian filter. These formalisms are able to accommodate more complex behavior, such as affect and habits, which might make them more competitive with RL models. I think it is entirely fair to say that these results demonstrate deviations from an idealized and strict Bayesian context; however, the equivalence here of Bayesian and normative is I think misleading or at least requires better justification/explanation. This is because a great deal of work has been done to show that Bayes optimal models can generate behavior or other outcomes that are clearly not optimal to an observer within a given context (consider hallucinations for example) but which make sense in the context of how the model is constructed as well as the priors and desired states the model is given.

As such, I would recommend that the language be adjusted to carefully define what is meant by normative and Bayesian and to recognize that work that is clearly Bayesian could potentially still be competitive with RL models if implemented to model this task. An even

better approach would be to directly use one of these more complex modelling approaches, such as active inference, as the comparator to the RL models, though I would understand if the authors would want this to be a subject for future work.

Abstract:

The abstract is lacking in some detail about the experiments done, but this may be a limitation of the required word count? If word count is not an issue, I would recommend adding details of the experiments done and the results. One comment is that there is an appeal to normative learning patterns, but this suggests that learning patterns have a fixed optimal nature, which may not be true in cases where the purpose of the learning (e.g. to confirm the feeling of safety of being in an in-group) may not be about learning accurately to maximize reward. This can be accommodated in a Bayesian framework by modelling priors and desired outcomes. As such the central premise that biased learning is inherently non-normative or non-Bayesian I think would require more justification. This is true in the introduction as well.

Introduction:

As noted above the conceptualization of Bayesian learning being equivalent to normative learning I think requires either further justification. Bayesian belief updating can be biased an non-optimal from an observer perspective, while being optimal within the agent doing the updating if the priors/desired outcomes are set up to advantage these "non-optimal" modes of decision making.

Results:

I wonder why the agent was presented before the choice - since the agent is only relevant to the feedback after the choice is made. I wonder if that might have induced any false association between the agent identity and the choice itself. This is by no means a critical point but would be interesting to get the authors' thoughts.

The finding that positive feedback increases learning is one that has been shown before and depends on valence, as the authors note. They expanded their reinforcement learning model to include valence; but they did not modify the Bayesian model in a similar manner. This lack of a valence or recency effect might also explain the failure of the Bayesian models in the preceding section where the contrast effect is discussed. It is not unreasonable to imagine that if humans do employ Bayesian reasoning that this reasoning system has had parameters tuned based on the real world, where recency of information does matter; affect has also been shown to be incorporable into Bayesian information processing (see the work by Hesp on affective charge and the large body of work by Ryan Smith). It may be that the Bayesian models chosen here require further complexity to capture the situation, just like some of the biases required updates to the RL models. This complexity, rather than being arbitrary, may be well justified by decision making in the real world.

The methods mention several symptom scales- it would be interesting to have the results of these and any interesting correlations noted. It is possible that some of individual variability here could be related to these symptoms, which could introduce precision parameter changes in a Bayesian context and things like reward sensitivity changes in an RL context.

Discussion:

(For discussion, not a specific comment on this paper): One wonders also about participant beliefs about the experiment or the intent of the experimenters. I have often had participants tell me they were trying to "figure out" a task or find patterns even when this was not part of the experiment. This is not specific to this paper, but it may be relevant in the future to try

and model participant beliefs about the experiment especially in the context of disinformation, when they might be primed to try and "figure things out".

As a general comment, in the active inference literature, there has been discussion of state-dependent actions, or "habits", which are learned in order to help agents more rapidly make decisions, based on previous learning. It is also possible that what is being observed is that these habits are at play, and that they represent the cognitive biases. This is likely especially true given, as the authors note, the high cognitive load of the task. It is true that this would mean that full-force Bayesian inference is not being used in each trial, or in each experience an agent might have in the world, but this is likely adaptive on the longer timescale of things, considering resource requirements. I think in this case you could argue that we have a departure from "normative" learning, but that is not necessarily a departure from any possible Bayesian framework, since these biases could potentially be modified by the agent or eschewed in favor of more expensive full-on Bayesian learning when warranted. Indeed in their discussion on the strategy of amplifying credible news sources to drown out low-credibility sources, the authors hint to the possibility of longer term strategies that may produce optimal outcomes in some contexts, but which were not necessarily appropriate to this task. As such, the performance on this task- and the consideration of true departure from Bayesian processing- should be considered in this wider context. Another thing to consider is that Bayesian inference is occurring, but that priors present going in produce the biases, or these biases arise from another source, for example factoring in epistemic value over rewards when the actual reward is not large. This again would be covered under an active inference approach, depending on how the priors are tuned. Indeed, given the benefit of social cohesion in an evolutionary perspective, some of these "biases" may be the result of adaptation. For example, it might be better to amplify people's good qualities and minimize their bad qualities in order to make it easier to interact with them; this entails a cost (in this case, not adequately learning from feedback and potentially losing out sometimes), but may fulfill a greater imperative (improved cooperation on things that matter). Given the right priors/desired states, this could still be a Bayes-optimal inference at a social level and as such may be ingrained as a habit which requires effort to break at the individual level during a task such as this.

The authors note that this task does not relate to "emotional engagement" or "deep, identity-related, issues". While I agree that this is likely mostly true, it is also possible that just being told one is being lied to might elicit an emotional response that could bias responses, even if this is a weak response.

Comments on revisions:

In their updated version the authors have made some edits to address my concerns regarding the framing of the 'normative' bayesian model, clarifying that they utilized a simple bayesian model which is intended to adhere in an idealized manner to the intended task structure, though further simulations would have been ideal.

The authors, however, did not take my recommendation to explore the symptoms in the symptom scales they collected as being a potential source of variability. They note that these were for hypothesis generation and were exploratory, fair enough, but this study is not small and there should have been sufficient sample size for a very reasonable analysis looking at symptom scores.

However, overall the toned down claims and clarifications of intent are adequate responses to my previous review.

https://doi.org/10.7554/eLife.106073.2.sa1

Reviewer #2 (Public review):

This important paper studies the problem of learning from feedback given by sources of varying credibility. The convincing combination of experiment and computational modeling helps to pin down properties of learning, while opening unresolved questions for future research.

Summary:

This paper studies the problem of learning from feedback given by sources of varying credibility. Two bandit-style experiments are conducted in which feedback is provided with uncertainty, but from known sources. Bayesian benchmarks are provided to assess normative facets of learning, and alternative credit assignment models are fit for comparison. Some aspects of normativity appear, in addition to possible deviations such as asymmetric updating from positive and negative outcomes.

Strengths:

The paper tackles an important topic, with a relatively clean cognitive perspective. The construction of the experiment enables the use of computational modeling. This helps to pinpoint quantitatively the properties of learning and formally evaluate their impact and importance. The analyses are generally sensible, and advanced parameter recovery analyses (including cross-fitting procedure) provide confidence in the model estimation and comparison. The authors have very thoroughly revised the paper in response to previous comments.

Weaknesses:

The authors acknowledge the potential for cognitive load and the interleaved task structure to play a meaningful role in the results, though leave this for future work. This is entirely reasonable, but remains a limitation in our ability to generalize the results. Broadly, some of the results obtain in cases where the extent of generalization is not always addressed and remains uncertain.

https://doi.org/10.7554/eLife.106073.2.sa2

Reviewer #3 (Public review):

Summary

This paper investigates how disinformation affects reward learning processes in the context of a two-armed bandit task, where feedback is provided by agents with varying reliability (with lying probability explicitly instructed). They find that people learn more from credible sources, but also deviate systematically from optimal Bayesian learning: They learned from uninformative random feedback, learned more from positive feedback, and updated too quickly from fully credible feedback (especially following low-credibility feedback). Overall, this study highlights how misinformation could distort basic reward learning processes, without appeal to higher order social constructs like identity.

Strengths

- The experimental design is simple and well-controlled; in particular, it isolates basic learning processes by abstracting away from social context
- Modeling and statistics meet or exceed standards of rigor

- Limitations are acknowledged where appropriate, especially those regarding external validity
- The comparison model, Bayes with biased credibility estimates, is strong; deviations are much more compelling than e.g. a purely optimal model
- The conclusions are of substantial interest from both a theoretical and applied perspective

Weaknesses

The authors have addressed most of my concerns with the initial submission. However, in my view, evidence for the conclusion that less credible feedback yields a stronger positivity bias remains weak. This is due to two issues.

Absolute or relative positivity bias?

The conclusion of greater positivity bias for lower credible feedback (Fig 5) hinges on the specific way in which positivity bias is defined. Specifically, we only see the effect when normalizing the difference in sensitivity to positive vs. negative feedback by the sum. I appreciate that the authors present both and add the caveat whenever they mention the conclusion. However, without an argument that the relative definition is more appropriate, the fact of the matter is that the evidence is equivocal.

There is also a good reason to think that the *absolute* definition is more appropriate. As expected, participants learn more from credible feedback. Thus, normalizing by average learning (as in the relative definition) amounts to dividing the absolute difference by increasingly large numbers for more credible feedback. If there is a fixed absolute positivity bias (or something that looks like it), the relative bias will necessarily be lower for more credible feedback. In fact, the authors own results demonstrate this phenomenon (see below). A reduction in relative bias thus provides weak evidence for the claim.

It is interesting that the discovery study shows evidence of a drop in absolute bias. However, for me, this just raises questions. Why is there a difference? Was one a just a fluke? If so, which one?

Positivity bias or perseveration?

Positivity bias and perseveration will both predict a stronger relationship between positive (vs. negative) feedback and future choice. They can thus be confused for each other when inferred from choice data. This potentially calls into question all the results on positivity bias.

The authors clearly identify this concern in the text and go to considerable lengths to rule it out. However, the new results (in revision 1) show that a perseveration-only model can in fact account for the qualitative pattern in the human data (the CA parameters). This contradicts the current conclusion:

Critically, however, these analyses also confirmed that perseveration cannot account for our main finding of increased positivity bias, relative to the overall extent of CA, for low-credibility feedback.

Figure 24c shows that the credibility-CA model does in fact show stronger positivity bias for less credible feedback. The model distribution for credibility 1 is visibly lower than for credibilities 0.5 and 0.75.

The authors need to be clear that it is the *magnitude* of the effect that the perseveration-only model cannot account for. Furthermore, they should additionally clarify that this is true only for models fit to data; it is possible that the credibility-CA model could capture the full size of

the effect with different parameters (which could fit best if the model was implemented slightly differently).

The authors could make the new analyses somewhat stronger by using parameters optimized to capture just the pattern in CA parameters (for example by MSE). This would show that the models are in principle incapable of capturing the effect. However, this would be a marginal improvement because the conclusion would still rest on a quantitative difference that depends on specific modeling assumptions.

New simulations clearly demonstrate the confound in relative bias

Figure 24 also speaks to the relative vs. absolute question. The model without positivity bias shows a slightly *stronger* absolute "positivity bias" for the most credible feedback, but a *weaker* relative bias. This is exactly in line with the logic laid out above. In standard bandit tasks, perseveration can be quite well-captured by a fixed absolute positivity bias, which is roughly what we see in the simulations (I'm not sure what to make of the slight increase; perhaps a useful lead for the authors). However, when we divide by average credit assignment, we now see a reduction. This clearly demonstrates that a reduction in relative bias can emerge without any true differences in positivity bias.

Given everything above, I think it is unlikely that the present data can provide even "solid" evidence for the claim that positivity bias is greater with less credible feedback. This confound could be quickly ruled out, however, by a study in which feedback is sometimes provided in the absence of a choice. This would empirically isolate positivity bias from choice-related effects, including perseveration.

https://doi.org/10.7554/eLife.106073.2.sa3

Author response:

The following is the authors' response to the original reviews

> ***Reviewer #1 (Public review):***
>
> *This is a well-designed and very interesting study examining the impact of imprecise feedback on outcomes in decision-making. I think this is an important addition to the literature, and the results here, which provide a computational account of several decision-making biases, are insightful and interesting.*

We thank the reviewer for highlighting the strengths of this work.

> *I do not believe I have substantive concerns related to the actual results presented; my concerns are more related to the framing of some of the work. My main concern is regarding the assertion that the results prove that non-normative and non-Bayesian learning is taking place. I agree with the authors that their results demonstrate that people will make decisions in ways that demonstrate deviations from what would be optimal for maximizing reward in their task under a strict application of Bayes' rule. I also agree that they have built reinforcement learning models that do a good job of accounting for the observed behavior. However, the Bayesian models included are rather simple, per the author's descriptions, applications of Bayes' rule with either fixed or learned credibility for the feedback agents. In contrast, several versions of the RL models are used, each modified to account for different possible biases. However, more complex Bayes-based models exist, notably active inference, but even the hierarchical Gaussian filter. These formalisms are able to accommodate more complex behavior, such as affect and habits, which might make them more competitive with RL models. I think it is entirely*

> *fair to say that these results demonstrate deviations from an idealized and strict Bayesian context; however, the equivalence here of Bayesian and normative is, I think, misleading or at least requires better justification/explanation. This is because a great deal of work has been done to show that Bayes optimal models can generate behavior or other outcomes that are clearly not optimal to an observer within a given context (consider hallucinations for example), but which make sense in the context of how the model is constructed as well as the priors and desired states the model is given.*
>
> *As such, I would recommend that the language be adjusted to carefully define what is meant by normative and Bayesian and to recognize that work that is clearly Bayesian could potentially still be competitive with RL models if implemented to model this task. An even better approach would be to directly use one of these more complex modelling approaches, such as active inference, as the comparator to the RL models, though I would understand if the authors would want this to be a subject for future work.*

We thank the reviewer for raising this crucial and insightful point regarding the framing of our results and the definitions of 'normative' and 'Bayesian' learning. Our primary aim in this work was to characterize specific behavioral signatures that demonstrate deviations from predictions generated by a strict, idealized Bayesian framework when learning from disinformation (which we term “biases”). We deliberately employed relatively simple Bayesian models as benchmarks to highlight these specific biases. We fully agree that more sophisticated Bayes-based models (as mentioned by the reviewer, or others) could potentially offer alternative mechanistic explanations for participant behavior. However, we currently do not have a strong notion about which Bayesian models can encompass our findings, and hence, we leave this important question for future work.

To enhance clarity within the current manuscript we now avoided the use of the term “normative” to refer to our Bayesian models, using the term “ideal” instead. We also define more clearly what exactly we mean by that notion when the idea model is described:

“This model is based on an idealized assumptions that during the feedback stage of each trial, the value of the chosen bandit is updated (based on feedback valence and credibility) according to Bayes rule reflecting perfect adherence to the instructed task structure (i.e., how true outcomes and feedback are generated).”

Moreover, we have added a few sentences in the discussion commenting on how more complex Bayesian models might account for our empirical findings:

“However, as hypothesized, when facing potential disinformation, we also find that individuals exhibit several important biases i.e., deviations from strictly idealized Bayesian strategies. Future studies should explore if and under what assumptions, about the task’s generative structure and/or learner’s priors and objectives, more complex Bayesian models (e.g., active inference (58)) might account for our empirical findings.”

> *Abstract:*
>
> *The abstract is lacking in some detail about the experiments done, but this may be a limitation of the required word count. If word count is not an issue, I would recommend adding details of the experiments done and the results.*

We thank the reviewer for their valuable suggestion. We have now included more details about the experiment in the abstract:

“In two experiments, participants completed a two-armed bandit task, where they repeatedly chose between two lotteries and received outcome-feedback from sources of varying

credibility, who occasionally disseminated disinformation by lying about true choice outcome (e.g., reporting non reward when a reward was truly earned or vice versa)."

> *One comment is that there is an appeal to normative learning patterns, but this suggests that learning patterns have a fixed optimal nature, which may not be true in cases where the purpose of the learning (e.g. to confirm the feeling of safety of being in an in-group) may not be about learning accurately to maximize reward. This can be accommodated in a Bayesian framework by modelling priors and desired outcomes. As such, the central premise that biased learning is inherently non-normative or non-Bayesian, I think, would require more justification. This is true in the introduction as well.*
>
> *Introduction:*
>
> *As noted above, the conceptualization of Bayesian learning being equivalent to normative learning, I think requires further justification. Bayesian belief updating can be biased and non-optimal from an observer perspective, while being optimal within the agent doing the updating if the priors/desired outcomes are set up to advantage these "non-optimal" modes of decision making.*

We appreciate the reviewer's thoughtful comment regarding the conceptualization of "normative" and "Bayesian" learning. We fully agree that the definition of "normative" is nuanced and can indeed depend on whether one considers reward-maximization or the underlying principles of belief updating. As explained above we now restrict our presentation to deviations from “ideal Bayes” learning patterns and we acknowledge the reviewer’s concern in a caveat in our discussion.

> *Results:*
>
> *I wonder why the agent was presented before the choice, since the agent is only relevant to the feedback after the choice is made. I wonder if that might have induced any false association between the agent identity and the choice itself. This is by no means a critical point, but it would be interesting to get the authors' thoughts.*

We thank the reviewer for raising this interesting point regarding the presentation of the agent before the choice. Our decision to present the agent at this stage was intentional, as our original experimental design aimed to explore the possible effects of "expected source credibility" on participants' choices (e.g., whether knowledge of feedback credibility will affect choice speed and accuracy). However, we found nothing that would be interesting to report.

> *The finding that positive feedback increases learning is one that has been shown before and depends on valence, as the authors note. They expanded their reinforcement learning model to include valence, but they did not modify the Bayesian model in a similar manner. This lack of a valence or recency effect might also explain the failure of the Bayesian models in the preceding section, where the contrast effect is discussed. It is not unreasonable to imagine that if humans do employ Bayesian reasoning that this reasoning system has had parameters tuned based on the real world, where recency of information does matter; affect has also been shown to be incorporable into Bayesian information processing (see the work by Hesp on affective charge and the large body of work by Ryan Smith). It may be that the Bayesian models chosen here require further complexity to capture the situation, just like some of the biases required updates to the RL models. This complexity, rather than being arbitrary, may be well justified by decision-making in the real world.*

Thanks for these additional important ideas which speak more to the notion that more complex Bayesian frameworks may account for biases we report.

> *The methods mention several symptom scales- it would be interesting to have the results of these and any interesting correlations noted. It is possible that some of the individual variability here could be related to these symptoms, which could introduce precision parameter changes in a Bayesian context and things like reward sensitivity changes in an RL context.*

We included these questionnaires for exploratory purposes, with the aim of generating informed hypotheses for future research into individual differences in learning. Given the preliminary nature of these analyses, we believe further research is required about this important topic.

> *Discussion:*
>
> *(For discussion, not a specific comment on this paper): One wonders also about participants' beliefs about the experiment or the intent of the experimenters. I have often had participants tell me they were trying to "figure out" a task or find patterns even when this was not part of the experiment. This is not specific to this paper, but it may be relevant in the future to try and model participant beliefs about the experiment especially in the context of disinformation, when they might be primed to try and "figure things out".*

We thank the reviewer for this important recommendation. We agree and this point is included in our caveat (cited above) that future research should address what assumptions about the generative task structure can allow Bayesian models to account for our empirical patterns.

> *As a general comment, in the active inference literature, there has been discussion of state-dependent actions, or "habits", which are learned in order to help agents more rapidly make decisions, based on previous learning. It is also possible that what is being observed is that these habits are at play, and that they represent the cognitive biases. This is likely especially true given, as the authors note, the high cognitive load of the task. It is true that this would mean that full-force Bayesian inference is not being used in each trial, or in each experience an agent might have in the world, but this is likely adaptive on the longer timescale of things, considering resource requirements. I think in this case you could argue that we have a departure from "normative" learning, but that is not necessarily a departure from any possible Bayesian framework, since these biases could potentially be modified by the agent or eschewed in favor of more expensive full-on Bayesian learning when warranted.*
> *Indeed, in their discussion on the strategy of amplifying credible news sources to drown out low-credibility sources, the authors hint at the possibility of longer-term strategies that may produce optimal outcomes in some contexts, but which were not necessarily appropriate to this task. As such, the performance on this task- and the consideration of true departure from Bayesian processing- should be considered in this wider context.*
>
> *Another thing to consider is that Bayesian inference is occurring, but that priors present going in produce the biases, or these biases arise from another source, for example, factoring in epistemic value over rewards when the actual reward is not large. This again would be covered under an active inference approach, depending on how the priors are tuned. Indeed, given the benefit of social cohesion in an evolutionary perspective, some of these "biases" may be the result of adaptation. For example, it might be better to amplify people's good qualities and minimize their bad qualities in order to make it*

> *easier to interact with them; this entails a cost (in this case, not adequately learning from feedback and potentially losing out sometimes), but may fulfill a greater imperative (improved cooperation on things that matter). Given the right priors/desired states, this could still be a Bayes-optimal inference at a social level and, as such, may be ingrained as a habit that requires effort to break at the individual level during a task such as this.*

We thank the reviewer for these insightful suggestions speaking further to the point about more complex Bayesian models.

> *The authors note that this task does not relate to "emotional engagement" or "deep, identity-related issues". While I agree that this is likely mostly true, it is also possible that just being told one is being lied to might elicit an emotional response that could bias responses, even if this is a weak response.*

We agree with the reviewer that a task involving performance-based bonuses, and particularly one where participants are explicitly told they are being lied to, might elicit weak emotional response. However, our primary point is that the degree of these responses is expected to be substantially weaker than those typically observed in the broader disinformation literature, which frequently deals with highly salient political, social, or identity-related topics that inherently carry strong emotional and personal ties for participants, leading to much more pronounced affective engagement and potential biases. Our task deliberately avoids such issues thus minimizing the potential for significant emotion-driven biases. We have toned down the discussion accordingly:

“This occurs even when the decision at hand entails minimal emotional engagement or pertinence to deep, identity-related, issues.”

> ***Reviewer #2 (Public review):***
>
> *This valuable paper studies the problem of learning from feedback given by sources of varying credibility. The solid combination of experiment and computational modeling helps to pin down properties of learning, although some ambiguity remains in the interpretation of results.*
>
> *Summary:*
>
> *This paper studies the problem of learning from feedback given by sources of varying credibility. Two banditstyle experiments are conducted in which feedback is provided with uncertainty, but from known sources. Bayesian benchmarks are provided to assess normative facets of learning, and alternative credit assignment models are fit for comparison. Some aspects of normativity appear, in addition to deviations such as asymmetric updating from positive and negative outcomes.*
>
> *Strengths:*
>
> *The paper tackles an important topic, with a relatively clean cognitive perspective. The construction of the experiment enables the use of computational modeling. This helps to pinpoint quantitatively the properties of learning and formally evaluate their impact and importance. The analyses are generally sensible, and parameter recovery analyses help to provide some confidence in the model estimation and comparison.*

We thank the reviewer for highlighting the strengths of this work.

> *Weaknesses:*

> *(1) The approach in the paper overlaps somewhat with various papers, such as Diaconescu et al. (2014) and Schulz et al. (forthcoming), which also consider the Bayesian problem of learning and applying source credibility, in terms of theory and experiment. The authors should discuss how these papers are complementary, to better provide an integrative picture for readers.*
>
> *Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., ... & Stephan, K. E. (2014). Inferring the intentions of others by hierarchical Bayesian learning. PLoS computational biology, 10(9), e1003810.*
>
> *Schulz, L., Schulz, E., Bhui, R., & Dayan, P. Mechanisms of Mistrust: A Bayesian Account of Misinformation Learning. https://doi.org/10.31234/osf.io/8egxh*

We thank the reviewers for pointing us to this relevant work. We have updated the introduction, mentioning these precedents in the literature and highlighting our specific contributions:

"To address these questions, we adopt a novel approach within the disinformation literature by exploiting a Reinforcement Learning (RL) experimental framework (36). While RL has guided disinformation research in recent years (37–41), our approach is novel in using one of its most popular tasks: the "bandit task"."

We also explain in the discussion how these papers relate to the current study:

"Unlike previous studies wherein participants had to infer source credibility from experience (30,37,72), we took an explicit-instruction approach, allowing us to precisely assess source-credibility impact on learning, without confounding it with errors in learning about the sources themselves. More broadly, our work connects with prior research on observational learning, which examined how individuals learn from the actions or advice of social partners (72–75). This body of work has demonstrated that individuals integrate learning from their private experiences with learning based on others' actions or advice—whether by inferring the value others attribute to different options or by mimicking their behavior (57,76). However, our task differs significantly from traditional observational learning. Firstly, our feedback agents interpret outcomes rather than demonstrating or recommending actions (30,37,72)."

> *(2) It isn't completely clear what the "cross-fitting" procedure accomplishes. Can this be discussed further?*

We thank the reviewer for requesting further clarification on the cross-fitting procedure. Our study utilizes two distinct model families: Bayesian models and CA models. The credit assignment parameters from the CA models can be treated as "data/behavioural features" corresponding to how choice feedback affects choice-propensities. The cross fitting-approach allows us in effect to examine whether these propensity features are predicted from our Bayesian models. To the extent they are not, we can conclude empirical behavior is "biased".

Thus, in our cross-fitting procedure we compare the CA model parameters extracted from participant data (empirical features) with those that would be expected if our Bayesian agents performed the task. Specifically, we first fit participant behavior with our Bayesian models, then simulate this model using the best-fitted parameters and fit those simulations with our CA models. This generates a set of CA parameters that would be predicted if participants behavior is reduced to a Bayesian account. By comparing these predicted Bayesian CA parameters with the actual CA parameters obtained from human participants, the cross-fitting procedure allows us to quantitatively demonstrate that the observed participant parameters are indeed statistically significant deviations from normative

Bayesian processing. This provides a robust validation that the biases we identify are not artifacts of the CA model's structure but true departures from normative learning.

We also note that Reviewer 3 suggested an intuitive way to think about the CA parameters—as analogous to logistic regression coefficients in a "sophisticated regression" of choice on (recencyweighted) choice-feedback. We find this suggestion potentially helpful for readers. Under this interpretation, the purpose of the cross-fitting method can be seen simply as estimating the regression coefficients that would be predicted by our Bayesian agents, and comparing those to the empirical coefficients.

In our manuscript we now explain this issues more clearly by explaining how our model is analogous to a logistic regression:

"The probability to choose a bandit (say A over B) in this family of models is a logistic function of the contrast choice-propensities between these two bandits. One interpretation of this model is as a "sophisticated" logistic regression, where the CA parameters take the role of "regression coefficients" corresponding to the change in log odds of repeating the just-taken action in future trials based on the feedback (+/- CA for positive or negative feedback, respectively; the model also includes gradual perseveration which allows for constant log-odd changes that are not affected by choice feedback) . The forgetting rate captures the extent to which the effect of each trial on future choices diminishes with time. The Q-values are thus exponentially decaying sums of logistic choice propensities based on the types of feedback a bandit received."

We also explain our cross-fitting procedure in more detail:

"To further characterise deviations between behaviour and our Bayesian learning models, we used a "crossfitting" method. Treating CA parameters as data-features of interest (i.e., feedback dependent changes in choice propensity), our goal was to examine if and how empirical features differ from features extracted from simulations of our Bayesian learning models. Towards that goal, we simulated synthetic data based on Bayesian agents (using participants' best fitting parameters), but fitted these data using the CA-models, obtaining what we term "Bayesian-CA parameters" (Fig. 2d; Methods). A comparison of these BayesianCA parameters, with empirical-CA parameters obtained by fitting CA models to empirical data, allowed us to uncover patterns consistent with, or deviating from, ideal-Bayesian value-based inference. Under the sophisticated logistic-regression interpretation of the CA-model family the cross-fitting method comprises a comparison between empirical regression coefficients (i.e., empirical CA parameters) and regression coefficient based on simulations of Bayesian models (Bayesian CA parameters)."

> *(3) The Credibility-CA model seems to fit the same as the free-credibility Bayesian model in the first experiment and barely better in the second experiment. Why not use a more standard model comparison metric like the Bayesian Information Criterion (BIC)? Even if there are advantages to the bootstrap method (which should be described if so), the BIC would help for comparability between papers.*

We thank the reviewer for this important comment regarding our model comparison approach. We acknowledge that classical information criteria like AIC and BIC are widely used in RL studies. However, we argue our method for model-comparison is superior.

We conducted a model recovery analysis demonstrating a significant limitation of using AIC or BIC for model-comparison in our data. Both these methods are strongly biased in favor of the Bayesian models. Our PBCM method, on the other hand, is both unbiased and more accurate. We believe this is because "off the shelf" methods like AIC and BIC rely on strong assumptions (such as asymptotic sample size and trial-independence) that are not necessarily met in our tasks (Data is finite; Trials in RL tasks depend on previous trials). PBCM avoids

such assumptions to obtain comparison criteria specifically tailored to the structure and size of our empirical data. We have now mentioned this fact in the results section of the main text:

“We considered using AIC and BIC, which apply “off-the shelf” penalties for model-complexity. However, these methods do not adapt to features like finite sample size (relying instead on asymptotic assumption) or temporal dependence (as is common in reinforcement learning experiments). In contrast, the parametric bootstrap cross-fitting method replaces these fixed penalties with empirical, data-driven criteria for modelselection. Indeed, model-recovery simulations confirmed that whereas AIC and BIC were heavily biased in favour of the Bayesian models, the bootstrap method provided excellent model-recovery (See Fig. S20).”

We have also included such model recovery in the SI document:

> *(4) As suggested in the discussion, the updating based on random feedback could be due to the interleaving of trials. If one is used to learning from the source on most trials, the occasional random trial may be hard to resist updating from. The exact interleaving structure should also be clarified (I assume different sources were shown for each bandit pair). This would also relate to work on RL and working memory: Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. European Journal of Neuroscience, 35(7), 10241035.*

We thank the reviewer for this point. The specific interleaved structure of the agents is described in the main text:

“Each agent provided feedback for 5 trials for each bandit pair (with the agent order interleaved within the bandit pair).”

As well as in the methods section:

“Feedback agents were randomly interleaved across trials subject to the constraint that each agent appeared on 5-trials for each bandit pair.”

We also thank the reviewer for mentioning the relevant work on working memory. We have now added it to our discussion point:

“In our main study, we show that participants revised their beliefs based on entirely non-credible feedback, whereas an ideal Bayesian strategy dictates such feedback should be ignored. This finding resonates with the “continued-influence effect” whereby misleading information continues to influence an individual's beliefs even after it has been retracted (59,60). One possible explanation is that some participants failed to infer that feedback from the 1-star agent was statistically void of information content, essentially random (e.g., the group-level credibility of this agent was estimated by our free-credibility Bayesian model as higher than 50%). Participants were instructed that this feedback would be “a lie” 50% of the time but were not explicitly told that this meant it was random and should therefore be disregarded. Notably, however, there was no corresponding evidence random feedback affected behaviour in our discovery study. It is possible that an individual’s ability to filter out random information might have been limited due to a high cognitive load induced by our main study task, which required participants to track the values of three bandit pairs and juggle between three interleaved feedback agents (whereas in our discovery study each experimental block featured a single bandit pair). Future studies should explore more systematically how the ability to filter random feedback depends on cognitive load (61).”

> *(5) Why does the choice-repetition regression include "only trials for which the last same-pair trial featured the 3-star agent and in which the context trial featured a different*

> *bandit pair"? This could be stated more plainly.*

We thank the reviewer for this question. When we previously submitted our manuscript, we thought that finding enhanced credit-assignment for fully credible feedback following potential disinformation from a different context would constitute a striking demonstration of our “contrast effect”. However, upon reexamining this finding we found out we had a coding error (affecting how trials were filtered). We have now rerun and corrected this analysis. We have assessed the contrast effect for both "same-context" trials (where the contextual trial featured the same bandit pair as the learning trial) and "different-context" trials (where the contextual trial featured a different bandit pair). Our re-analysis reveals a selective significant contrast effect in the samecontext condition, but no significant effect in the different-context condition. We have updated the main text to reflect these corrected findings and provide a clearer explanation of the analysis:

“A comparison of empirical and Bayesian credit-assignment parameters revealed a further deviation from ideal Bayesian learning: participants showed an exaggerated credit-assignment for the 3-star agent compared with Bayesian models [Wilcoxon signed-rank test, instructed-credibility Bayesian model (median difference=0.74, z=11.14); free-credibility Bayesian model (median difference=0.62, z=10.71), all p’s<0.001] (Fig. 3a). One explanation for enhanced learning for the 3-star agents is a contrast effect, whereby credible information looms larger against a backdrop of non-credible information. To test this hypothesis, we examined whether the impact of feedback from the 3-star agent is modulated by the credibility of the agent in the trial immediately preceding it. More specifically, we reasoned that the impact of a 3-star agent would be amplified by a “low credibility context” (i.e., when it is preceded by a low credibility trial). In a binomial mixed effects model, we regressed choice-repetition on feedback valence from the last trial featuring the same bandit pair (i.e., the learning trial) and the feedback agent on the trial immediately preceding that last trial (i.e., the contextual credibility; see Methods for model-specification). This analysis included only learning trials featuring the 3-star agent, and context trials featuring the same bandit pair as the learning trial (Fig. 4a). We found that feedback valence interacted with contextual credibility ($F(2,2086)=11.47$, $p<0.001$) such that the feedback-effect (from the 3-star agent) decreased as a function of the preceding context-credibility (3-star context vs. 2-star context: $b= -0.29$, $F(1,2086)=4.06$, $p=0.044$; 2star context vs. 1-star context: $b=-0.41$, $t(2086)=-2.94$, $p=0.003$; and 3-star context vs. 1-star context: $b=0.69$, $t(2086)=-4.74$, $p<0.001$) (Fig. 4b). This contrast effect was not predicted by simulations of our main models of interest (Fig. 4c). No effect was found when focussing on contextual trials featuring a bandit pair different than the one in the learning trial (see SI 3.5). Thus, these results support an interpretation that credible feedback exerts a greater impact on participants’ learning when it follows non-credible feedback, in the same learning context.”

We have modified the discussion accordingly as well:

“A striking finding in our study was that for a fully credible feedback agent, credit assignment was exaggerated (i.e., higher than predicted by our Bayesian models). Furthermore, the effect of fully credible feedback on choice was further boosted when it was preceded by a low-credibility context related to current learning. We interpret this in terms of a “contrast effect”, whereby veridical information looms larger against a backdrop of disinformation (21). One upshot is that exaggerated learning might entail a risk of jumping to premature conclusions based on limited credible evidence (e.g., a strong conclusion that a vaccine is produces significant side-effect risks based on weak credible information, following non-credible information about the same vaccine). An intriguing possibility, that could be tested in future studies, is that participants strategically amplify the extent of learning from credible feedback to dilute the impact of learning from noncredible feedback. For example, a person scrolling through a social media feed, encountering copious amounts of disinformation, might amplify the weight they assign to credible feedback in order to dilute effects of ‘fake

news'. Ironically, these results also suggest that public campaigns might be more effective when embedding their messages in low-credibility contexts , which may boost their impact."

And we have included some additional analyses in the SI document:

"3.5 Contrast effects for contexts featuring a different bandit

Given that we observed a contrast effect when both the learning and the immediately preceding "context trial" involved the same pair of bandits, we next investigated whether this effect persisted when the context trial featured a different bandit pair – a situation where the context would be irrelevant to the current learning. Again, we used in a binomial mixed effects model, regressing choice-repetition on feedback valence in the learning trial and the feedback agent in the context trial. This analysis included only learning trials featuring the 3-star agent, and context trials featuring a different bandit pair than the learning trial (Fig. S22a). We found no significant evidence of an interaction between feedback valence and contextual credibility ($F(2,2364)=0.21$, $p=0.81$) (Fig. S22b). This null result was consistent with the range of outcomes predicted by our main computational models (Fig. S22c).

We aimed to formally compare the influence of two types of contextual trials: those featuring the same bandit pair as the learning trial versus those featuring a different pair. To achieve this, we extended our mixedeffects model by incorporating a new predictor variable, "CONTEXT_TYPE" which coded whether the contextual trial involved the same bandit pair (coded as -0.5) or a different bandit pair (+0.5) compared to the learning trial. The Wilkinson notation for this expanded mixed-effects model is:

$$REPEAT \sim CONTEXT\_TYPE * FEEDBACK * (CONTEXT_{2\text{-star}} + CONTEXT_{3\text{-star}}) + BETTER + (1 \mid participant)$$

This expanded model revealed a significant three-way interaction between feedback valence, contextual credibility, and context type ($F(2,4451) = 7.71$, $p<0.001$). Interpreting this interaction, we found a 2-way interaction between context-source and feedback valence when the context was the same ($F(2,4451) = 12.03$, $p<0.001$), but not when context was different ($F(2,4451) = 0.23$, $p = 0.79$). Further interpreting the double feedback-valence * context-source interaction (for the same context) we obtained the same conclusions as reported in the main text."

> *(6) Why apply the "Truth-CA" model and not the Bayesian variant that it was motivated by?*

Thanks for this very useful suggestion. We are unsure if we fully understand the question. The Truth-CA model was not motivated by a new Bayesian model. Our Bayesian models were simply used to make the point that participants may partially discriminate between truthful and untruthful feedback (for a given source). This led to the idea that perhaps more credit is assigned for truth (than lie) trials, which is what we found using our Truth-CA model. Note we show that our Bayesian models cannot account for this modulation.

We have now improved our "Truth-CA" model. Previously, our Truth-CA model considered whether feedback on each trial was true or not based on realized latent true outcomes. However, it is possible that the very same feedback would have had an opposite truth-status if the latent true outcome was different (recall true outcomes are stochastic). This injects noise into the trial classification in our previous model. To avoid this, in our new model feedback is modulated by the probability the reported feedback is true (marginalized over stochasticity of true outcome).

We have described this new model in the methods section:

"Additionally, we formulated a "Truth-CA" model, which worked as our Credibility-CA model, but incorporated a free truth-bonus parameter (TB). This parameter modulates the extent of credit assignment for each agent based on the posterior probability of feedback being true (given the credibility of the feedback agent, and the true reward probability of the chosen bandit). The chosen bandit was updated as follows:

$$Q \leftarrow (1 - f_Q) * Q + [CA(agent) + TB * (P(truth) - 0.5)] * F$$

where P(truth) is the posterior probability of the feedback being true in the current trial (for exact calculation of P(truth) see "Methods: Bayesian estimation of posterior belief that feedback is true")."

All relevant results have been updated accordingly in the main text:

"To formally address whether feedback truthfulness modulates credit assignment, we fitted a new variant of the CA model (the "Truth-CA" model) to the data. This variant works as our Credibility-CA model but incorporated a truth-bonus parameter (TB) which increases the degree of credit assignment for feedback as a function of the experimenter-determined likelihood the feedback is true (which is read from the curves in Fig 6a when x is taken to be the true probability the bandit is rewarding). Specifically, after receiving feedback, the Q-value of the chosen option is updated according to the following rule: $Q \leftarrow (1 - f_Q) * Q + [CA(agent) + TB * (P(truth) - 0.5)] * F$ where TB is the free parameter representing the truth bonus, and $P(truth)$ is the probability the received feedback being true (from the experimenter's perspective). We acknowledge that this model falls short of providing a mechanistically plausible description of the credit assignment process, because participants have no access to the experimenter's truthfulness likelihoods (as the true bandit reward probabilities are unknown to them). Nonetheless, we use this 'oracle model' as a measurement tool to glean rough estimates for the extent to which credit assignment Is boosted as a function of its truthfulness likelihood. Fitting this Truth-CA model to participants' behaviour revealed a significant positive truth-bonus (mean=0.21, t(203)=3.12, p=0.002), suggesting that participants indeed assign greater weight to feedback that is likely to be true (Fig. 6c; see SI 3.3.1 for detailed ML parameter results). Notably, simulations using our other models (Methods) consistently predicted smaller truth biases (compared to the empirical bias) (Fig. 6d). Moreover, truth bias was still detected even in a more flexible model that allowed for both a positivity bias and truth-bias (see SI 3.7). The upshot is that participants are biased to assign higher credit based on feedback that is more likely to be true in a manner that is inconsistent with out Bayesian models and above and beyond the previously identified positivity biases."

Finally, the Supplementary Information for the discovery study has also been revised to feature this analysis:

"We next assessed whether participants infer whether the feedback they received on each trial was true or false and adjust their credit assignment based on this inference. We again used the "Truth-CA" model to obtain estimates for the truth bonus (TB), the increase in credit assignment as a function of the posterior probability of feedback being true. As in our main study, the fitted truth bias parameter was significantly positive, indicating that participants assign greater weight to feedback they believe is likely to be true (Fig, S4a; see SI 3.3.1 for detailed ML parameter results). Strikingly, model-simulations (Methods) predicted a lower truth bonus than the one observed in participants (Fig. S4b)."

> *(7) "Overall, the results from this study support the exact same conclusions (See SI section 1.2) but with one difference. In the discovery study, we found no evidence for learning based on 50%-credibility feedback when examining either the feedback effect on choice repetition or CA in the credibility-CA model (SI 1.2.3)" - this seems like a very salient*

> *difference, when the paper reports the feedback effect as a primary finding of interest, though I understand there remains a valence-based difference.*

We agree with the reviewer and thank them for this suggestion. We now state explicitly throughout the manuscript that this finding was obtained only in one of our two studies. In the section “Discovery study” of the results we state explicitly this finding was not found in the discovery study:

“However, we found no evidence for learning based on 50%-credibility feedback when examining either the feedback effect on choice repetition or CA in the credibility-CA model (SI 1.2.3).”

We also note that related to another concern from R3 (that perseveration may masquerade as positivity bias) we conducted additional analyses (detailed in SI 3.6.2). These analyses revealed that the observed positivity bias for the 1-star agent in the discovery study falls within the range predicted by simple choice-perseveration. Consequently, we have removed the suggestion that participants still learn from the random agent in the discovery study. Furthermore, we have modified the discussion section to include a possible explanation for this discrepancy between the two studies:

“Notably, however, there was no corresponding evidence random feedback affected behaviour in our discovery study. It is possible that an individual’s ability to filter out random information might have been limited due to a high cognitive load induced by our main study task, which required participants to track the values of three bandit pairs and juggle between three interleaved feedback agents (whereas in our discovery study each experimental block featured a single bandit pair). Future studies should explore more systematically how the ability to filter random feedback depends on cognitive load (61).”

> *(8) "Participants were instructed that this feedback would be "a lie 50% of the time but were not explicitly told that this meant it was random and should therefore be disregarded." - I agree that this is a possible explanation for updating from the random source. It is a meaningful caveat.*

Thank you for this thought. While this can be seen as a caveat—since we don’t know what would have happened with explicit instructions—we also believe it is interesting from another perspective. In many real-life situations, individuals may have all the necessary information to infer that the feedback they receive is uninformative, yet still fail to do so, especially when they are not explicitly told to ignore it.

In future work, we plan to examine how behaviour changes when participants are given more explicit instructions—for example, that the 50%-credibility agent provides purely random feedback.

> *(9) "Future studies should investigate conditions that enhance an ability to discard disinformation, such as providing explicit instructions to ignore misleading feedback, manipulations that increase the time available for evaluating information, or interventions that strengthen source memory." - there is work on some of this in the misinformation literature that should be cited, such as the "continued influence effect". For example: Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. Journal of experimental psychology: Learning, memory, and cognition, 20(6), 1420.*

We thank the reviewer for pointing us towards the relevant literature. We have now included citations about the “continued influence effect” of misinformation in the discussion:

"In our main study, we show that participants revised their beliefs based on entirely non-credible feedback, whereas an ideal Bayesian strategy dictates such feedback should be ignored. This finding resonates with the "continued-influence effect" whereby misleading information continues to influence an individual's beliefs even after it has been retracted (59,60)."

> *(10) Are the authors arguing that choice-confirmation bias may be at play? Work on choice-confirmation bias generally includes counterfactual feedback, which is not present here.*

We agree with the reviewer that a definitive test for choice-confirmation bias typically requires counterfactual feedback, which is not present in our current task. In our discussion, we indeed suggest that the positivity bias we observe may stem from a form of choice-confirmation, drawing on the extensive literature on this bias in reinforcement learning (Lefebvre et al., 2017; Palminteri et al., 2017; Palminteri & Lebreton, 2022). However, we fully acknowledge that this link is a hypothesis and that explicitly testing for choice-confirmation bias would necessitate a future study specifically incorporating counterfactual feedback. We have included a clarification of this point in the discussion:

"Previous reinforcement learning studies, report greater credit-assignment based on positive compared to negative feedback, albeit only in the context of veridical feedback (43,44,62). Here, supporting our a-priori hypothesis we show that this positivity bias is amplified for information of low and intermediate credibility (in absolute terms in the discovery study, and relative to the overall extent of CA in both studies) . Of note, previous literature has interpreted enhanced learning for positive outcomes in reinforcement learning as indicative of a confirmation bias (42,44). For example, positive feedback may confirm, to a greater extent than negative feedback one's choice as superior (e.g., "I chose the better of the two options"). Leveraging the framework of motivated cognition (35), we posited that feedback of uncertain veracity (e.g., low credibility) amplifies this bias by incentivising individuals to self-servingly accept positive feedback as true (because it confers positive, desirable outcomes), and explain away undesirable, choice-disconfirming, negative feedback as false. This could imply an amplified confirmation bias on social media, where content from sources of uncertain credibility, such as unknown or unverified users, is more easily interpreted in a self-serving manner, disproportionately reinforcing existing beliefs (63). In turn, this could contribute to an exacerbation of the negative social outcomes previously linked to confirmation bias such as polarization (64,65), the formation of 'echo chambers' (19), and the persistence of misbelief regarding contemporary issues of importance such as vaccination (66,67) and climate change (68–71). We note however, that further studies are required to determine whether positivity bias in our task is indeed a form of confirmation bias."

> ***Reviewer #3 (Public review):***
>
> *Summary*
>
> *This paper investigates how disinformation affects reward learning processes in the context of a two-armed bandit task, where feedback is provided by agents with varying reliability (with lying probability explicitly instructed). They find that people learn more from credible sources, but also deviate systematically from optimal Bayesian learning: They learned from uninformative random feedback, learned more from positive feedback, and updated too quickly from fully credible feedback (especially following low-credibility feedback). Overall, this study highlights how misinformation could distort basic reward learning processes, without appeal to higher-order social constructs like identity.*
>
> *Strengths*

> *(1) The experimental design is simple and well-controlled; in particular, it isolates basic learning processes by abstracting away from social context.*
>
> *(2) Modeling and statistics meet or exceed the standards of rigor.*
>
> *(3) Limitations are acknowledged where appropriate, especially those regarding external validity.*
>
> *(4) The comparison model, Bayes with biased credibility estimates, is strong; deviations are much more compelling than e.g., a purely optimal model.*
>
> *(5) The conclusions are interesting, in particular the finding that positivity bias is stronger when learning from less reliable feedback (although I am somewhat uncertain about the validity of this conclusion)*

We deeply thank the reviewer for highlighting the strengths of this work.

> *Weaknesses*
>
> *(1) Absolute or relative positivity bias?*
>
> *In my view, the biggest weakness in the paper is that the conclusion of greater positivity bias for lower credible feedback (Figure 5) hinges on the specific way in which positivity bias is defined. Specifically, we only see the effect when normalizing the difference in sensitivity to positive vs. negative feedback by the sum. I appreciate that the authors present both and add the caveat whenever they mention the conclusion (with the crucial exception of the abstract). However, what we really need here is an argument that the relative definition is the* right *way to define asymmetry....*
>
> *Unfortunately, my intuition is that the absolute difference is a better measure. I understand that the relative version is common in the RL literature; however previous studies have used standard TD models, whereas the current model updates based on the raw reward. The role of the CA parameter is thus importantly different from a traditional learning rate - in particular, it's more like a logistic regression coefficient (as described below) because it scales the feedback but* not *the decay. Under this interpretation, a difference in positivity bias across credibility conditions corresponds to a three-way interaction between the exponentially weighted sum of previous feedback of a given type (e.g., positive from the 75% credible agent), feedback positivity, and condition (dummy coded). This interaction corresponds to the nonnormalized, absolute difference.*
>
> *Importantly, I'm not terribly confident in this argument, but it does suggest that we need a compelling argument for the relative definition.*

We thank the reviewer for raising this important point about the definition of positivity bias, and for their thoughtful discussion on the absolute versus relative measures. We believe that the relative valence bias offers a distinct and valuable perspective on positivity bias. Conceptually, this measure describes positivity bias in a manner akin to a “percentage difference” relative to the overall level of learning which allows us to control for the overall decreases in the overall amount of credit assignment as feedback becomes less credible. We are unsure if one measure is better or more correct than the other and we believe that reporting both measures enriches the understanding of positivity bias and allows for a more comprehensive characterization of this phenomenon (as long as these measures are interpreted carefully). We have stated the significance of the relative measure in the results section:

"Following previous research, we quantified positivity bias in 2 ways: 1) as the absolute difference between credit-assignment based on positive or negative feedback, and 2) as the same difference but relative to the overall extent of learning. We note that the second, relative, definition, is more akin to "percentage change" measurements providing a control for the overall lower levels of credit-assignment for less credible agent."

We also wish to point out that in our discovery study we had some evidence for amplification of positivity bias in absolute sense.

> *(2) Positivity bias or perseveration?*
>
> *A key challenge in interpreting many of the results is dissociating perseveration from other learning biases. In particular, a positivity bias (Figure 5) and perseveration will both predict a stronger correlation between positive feedback and future choice. Crucially, the authors do include a perseveration term, so one would hope that perseveration effects have been controlled for and that the CA parameters reflect true positivity biases. However, with finite data, we cannot be sure that the variance will be correctly allocated to each parameter (c.f. collinearity in regressions). The fact that CA- is fit to be negative for many participants (a pattern shown more strongly in the discovery study) is suggestive that this might be happening. A priori, the idea that you would ever increase your value estimate after negative feedback is highly implausible, which suggests that the parameter might be capturing variance besides that it is intended to capture.*
>
> *The best way to resolve this uncertainty would involve running a new study in which feedback was sometimes provided in the absence of a choice - this would isolate positivity bias. Short of that, perhaps one could fit a version of the Bayesian model that also includes perseveration. If the authors can show that this model cannot capture the pattern in Figure 5, that would be fairly convincing.*

We thank the reviewer for this very insightful and crucial point regarding the potential confound between positivity bias and perseveration. We entirely agree that distinguishing these effects can be challenging. To rigorously address this concern and ascertain that our observed positivity bias, particularly its inflation for low-credibility feedback, is not merely an artifact of perseveration, we conducted additional analyses as suggested.

First, following the reviewer's suggestion we simulated our Bayesian models, including a perseveration term, for both our main and discovery studies. Crucially, none of these simulations predicted the specific pattern of inflated positivity bias for low-credibility feedback that we identified in participants.

Additionally, taking a "devil's advocate" approach, we tested whether our credibility-CA model (which includes perseveration but not a feedback valence bias) can predict our positivity bias findings. Thus, we simulated 100 datasets using our Credibility-CA model (based on empirical best-fitting parameters). We then fitted each of these simulated datasets using our CredibilityValence CA model. By examining the distribution of results across these synthetic datasets fits and comparing them to the actual results from participants, we found that while perseveration could indeed lead (as the reviewer suspected) to an artifactual positivity bias, it could not predict the magnitude of the observed inflation of positivity bias for low-credibility feedback (whether measured in absolute or relative terms).

Based on these comprehensive analyses, we are confident that our main results concerning the modulation of a valence bias as a function of source-credibility cannot be accounted by simple choice-perseveration. We have briefly explained these analyses in the main results section:

"Previous research has suggested that positivity bias may spuriously arise from pure choice-perseveration (i.e., a tendency to repeat previous choices regardless of outcome) (49,50). While our models included a perseveration-component, this control may not be preferent. Therefore, in additional control analyses, we generated synthetic datasets using models including choice-perseveration but devoid of feedback-valence bias, and fitted them with our credibility-valence model (see SI 3.6.1). These analyses confirmed that perseveration can masquerade as an apparent positivity bias. Critically, however, these analyses also confirmed that perseveration cannot account for our main finding of increased positivity bias, relative to the overall extent of CA, for low-credibility feedback."

Additionally, we have added a detailed description of these additional analyses and their findings to the Supplementary Information document:

"3.6 Positivity bias results cannot be explained by a pure perseveration

3.6.1 Main study

Previous research has suggested it may be challenging to dissociate between a feedback-valence positivity bias and perseveration (i.e., a tendency to repeat previous choices regardless of outcome). While our Credit Assignment (CA) models already include a perseveration mechanism to account for this, this control may not be perfect. We thus conducted several tests to examine if our positivity-bias related results could be accounted for by perseveration.

First we examined whether our Bayesian-models, augmented by a perseveration mechanism (as in our CA model) can generate predictions similar to our empirical results. We repeated our cross-fitting procedure to these extended Bayesian models. To briefly recap, this involved fitting participant behavior with them, generating synthetic datasets based on the resulting maximum likelihood (ML) parameters, and then fitting these simulated datasets with our Credibility-Valence CA model (which is designed to detect positivity bias). This test revealed that adding perseveration to our Bayesian models did not predict a positivity bias in learning. In absolute terms there was a small negativity bias (instructed-credibility Bayesian: $b=-0.19$, $F(1,1218)=17.78$, $p<0.001$, Fig. S23a-b; free-credibility Bayesian: $b=-0.17$, $F(1,1218)=13.74$, $p<0.001$, Fig. S23d-e). In relative terms we detected no valence related bias (instructed-credibility Bayesian: $b=-0.034$, $F(1,609)=0.45$, $p=0.50$, Fig. S22c; free-credibility Bayesian: $b=-0.04$, $F(1,609)=0.51$, $p=0.47$, Fig. S23f). More critically, these simulations also did not predict a change in the level of positivity bias as a function of feedback credibility, neither at an absolute level (instructed-credibility Bayesian: $F(2,1218)=0.024$, $p=0.98$, Fig. S23b; free-credibility Bayesian: $F(2,1218)=0.008$, $p=0.99$, Fig. S23e), nor at a relative level (instructedcredibility Bayesian: $F(2,609)=1.57$, $p=0.21$, Fig. S23c; free-credibility Bayesian: $F(2,609)=0.13$, $p=0.88$, Fig. S23f). The upshot is that our positivity-bias findings cannot be accounted for by our Bayesian models even when these are augmented with perseveration.

However, it is still possible that empirical CA parameters from our credibility-valence model (reported in main text Fig. 5) were distorted, absorbing variance from a perseveration. To address this, we took a "devil's advocate" approach testing the assumption that CA parameters are not truly affected by feedback valance and that there is only perseveration in our data. Towards that goal, we simulated data using our CredibilityCA model (which includes perseveration but does not contain a valence bias in its learning mechanism) and then fitted these synthetic datasets using our Credibility-Valence CA model to see if the observed positivity bias could be explained by perseveration alone. Specifically, we generated 101 "group-level" synthetic datasets (each including one simulation for each participant, based on their empirical ML parameters), and fitted each dataset with our Credibility-Valence CA model. We then analysed the resulting ML parameters in each dataset using the same mixed-effects models as described in the main text, examining the distribution of effects of

interest across these simulated datasets. Comparing these simulation results to the data from participants revealed a nuanced picture. While the positivity bias observed in participants is within the range predicted by a pure perseveration account when measured in absolute terms (Fig. S24a), it is much higher than predicted by pure perseveration when measured relative to the overall level of learning (Fig. S24c). More importantly, the inflation in positivity bias for lower credibility feedback is substantially higher in participants than what would be predicted by a pure perseveration account, a finding that holds true for both absolute (Fig. S24b) and relative (Fig. S24d) measures."

"3.6.2 Discovery study

We then replicated these analyses in our discovery study to confirm our findings. We again checked whether extended versions of the Bayesian models (including perseveration) predicted the positivity bias results observed. Our cross-fitting procedure showed that the instructed-credibility Bayesian model with perseveration did predict a positivity bias for all credibility levels in this discovery study, both when measured in absolute terms [50% credibility ($b=1.74, t(824)=6.15$), 70% credibility ($b=2.00, F(1,824)=49.98$), 85% credibility ($b=1.81, F(1,824)=40.78$), 100% credibility ($b=2.42, F(1,824)=72.50$), all p's<0.001], and in relative terms [50% credibility ($b=0.25, t(412)=3.44$), 70% credibility ($b=0.31, F(1,412)=17.72$), 85% credibility ($b=0.34, F(1,412)=21.06$), 100% credibility ($b=0.42, F(1,412)=31.24$), all p's<0.001]. However, importantly, these simulations did not predict a change in the level of positivity bias as a function of feedback credibility, neither at an absolute level ($F(3,412)=1.43, p=0.24$), nor at a relative level ($F(3,412)=2.06, p=0.13$) (Fig. S25a-c). In contrast, simulations of the free-credibility Bayesian model (with perseveration) predicted a slight negativity bias when measured in absolute terms ($b=-0.35, F(1,824)=5.14, p=0.024$), and no valence bias when measured relative to the overall degree of learning ($b=0.05, F(1,412)=0.55, p=0.46$). Crucially, this model also did not predict a change in the level of positivity bias as a function of feedback credibility, neither at an absolute level ($F(3,824)=0.27, p=0.77$), nor at a relative level ($F(3,412)=0.76, p=0.47$) (Fig. S25d-f).

As in our main study, we next assessed whether our Credibility-CA model (which includes perseveration but no valence bias) predicted the positivity bias results observed in participants in the discovery study. This analysis revealed that the average positivity bias in participants is higher than predicted by a pure perseveration account, both when measured in absolute terms (Fig. S26a) and in relative terms (Fig. S26c). Specifically, only the aVBI for the 70% credibility agent was above what a perseveration account would predict, while the rVBI for all agents except the completely credible one exceeded that threshold. Furthermore, the inflation in positivity bias for lower credibility feedback (compared to the 100% credibility agent) is significantly higher in participants than would be predicted by a pure perseveration account, in both absolute (Fig. S26b) and relative (Fig. S26d) terms.

Together, these results show that the general positivity bias observed in participants could be predicted by an instructed-credibility Bayesian model with perseveration, or by a CA model with perseveration. Moreover, we find that these two models can predict a positivity bias for the 50% credibility agent, raising a concern that our positivity bias findings for this source may be an artefact of not-fully controlled for perseveration. However, the credibility modulation of this positivity bias, where the bias is amplified for lower credibility feedback, is consistently not predicted by perseveration alone, regardless of whether perseveration is incorporated into a Bayesian or a CA model. This finding suggests that participants are genuinely modulating their learning based on feedback credibility, and that this modulation is not merely an artifact of choice perseveration."

> *(3) Veracity detection or positivity bias?*
>
> *The "True feedback elicits greater learning" effect (Figure 6) may be simply a re-description of the positivity bias shown in Figure 5. This figure shows that people have*

> *higher CA for trials where the feedback was in fact accurate. But assuming that people tend to choose more rewarding options, true-feedback cases will tend to also be positive-feedback cases. Accordingly, a positivity bias would yield this effect, even if people are not at all sensitive to trial-level feedback veracity. Of course, the reverse logic also applies, such that the "positivity bias" could actually reflect discounting of feedback that is less likely to be true. This idea has been proposed before as an explanation for confirmation bias (see Pilgrim et al, 2024 https://doi.org/10.1016/j.cognition.2023.105693and much previous work cited therein). The authors should discuss the ambiguity between the "positivity bias" and "true feedback" effects within the context of this literature....*

Before addressing these excellent comments, we first note that we have now improved our "TruthCA" model. Previously, our Truth-CA model considered whether feedback on each trial was true or not based on realized latent true outcomes. However, it is possible that the very same feedback would have had an opposite truth-status if the latent true outcome was different (recall true outcomes are stochastic). This injects noise into the trial classification in our former model. To avoid this, in our new model feedback is modulated by the probability the reported feedback is true (marginalized over stochasticity of true outcome). Please note in our responses below that we conducted extensive analysis to confirm that positivity bias doesn't in fact predict the truthbias we detect using our truth biased model

We have described this new model in the methods section:

"Additionally, we formulated a "Truth-CA" model, which worked as our Credibility-CA model, but incorporated a free truth-bonus parameter (TB). This parameter modulates the extent of credit assignment for each agent based on the posterior probability of feedback being true (given the credibility of the feedback agent, and the true reward probability of the chosen bandit). The chosen bandit was updated as follows:

$$Q \leftarrow (1 - f_Q) * Q + [CA(agent) + TB * (P(truth) - 0.5)] * F$$

where P(truth) is the posterior probability of the feedback being true in the current trial (for exact calculation of P(truth) see "Methods: Bayesian estimation of posterior belief that feedback is true")."

All relevant results have been updated accordingly in the main text:

To formally address whether feedback truthfulness modulates credit assignment, we fitted a new variant of the CA model (the "Truth-CA" model) to the data. This variant works as our Credibility-CA model, but incorporated a truth-bonus parameter (TB) which increases the degree of credit assignment for feedback as a function of the experimenter-determined likelihood the feedback is true (which is read from the curves in Fig 6a when x is taken to be the true probability the bandit is rewarding). Specifically, after receiving feedback, the Q-value of the chosen option is updated according to the following rule:

$$Q \leftarrow (1 - f_Q) * Q + [CA(agent) + TB * (P(truth) - 0.5)] * F$$

where TB is the free parameter representing the truth bonus, and $P(truth)$ is the probability the received feedback being true (from the experimenter's perspective). We acknowledge that this model falls short of providing a mechanistically plausible description of the credit assignment process, because participants have no access to the experimenter's truthfulness likelihoods (as the true bandit reward probabilities are unknown to them). Nonetheless, we use this 'oracle model' as a measurement tool to glean rough estimates for the extent to which credit assignment Is boosted as a function of its truthfulness likelihood.

Fitting this Truth-CA model to participants' behaviour revealed a significant positive truth-bonus (mean=0.21, t(203)=3.12, p=0.002), suggesting that participants indeed assign greater

weight to feedback that is likely to be true (Fig. 6c; see SI 3.3.1 for detailed ML parameter results). Notably, simulations using our other models (Methods) consistently predicted smaller truth biases (compared to the empirical bias) (Fig. 6d). Moreover, truth bias was still detected even in a more flexible model that allowed for both a positivity bias and truth-bias (see SI 3.7). The upshot is that participants are biased to assign higher credit based on feedback that is more likely to be true in a manner that is inconsistent with out Bayesian models and above and beyond the previously identified positivity biases."

Finally, the Supplementary Information for the discovery study has also been revised to feature this analysis:

"We next assessed whether participants infer whether the feedback they received on each trial was true or false and adjust their credit assignment based on this inference. We again used the "Truth-CA" model to obtain estimates for the truth bonus (TB), the increase in credit assignment as a function of the posterior probability of feedback being true. As in our main study, the fitted truth bias parameter was significantly positive, indicating that participants assign greater weight to feedback they believe is likely to be true (Fig, S4a; see SI 3.3.1 for detailed ML parameter results). Strikingly, model-simulations (Methods) predicted a lower truth bonus than the one observed in participants (Fig. S4b)."

Additionally, we thank the reviewer for pointing us to the relevant work by Pilgrim et al. (2024). We agree that the relationship between "true feedback" and "positivity bias" effects is nuanced, and their potential overlap warrants careful consideration. Note our analyses suggest that this is not solely the case. Firstly, simulations of our Credibility-Valence CA model predict only a small "truth bonus" effect, which is notably smaller than what we observed in participants. Secondly, we formulated an extension of our "Truth-CA" model that includes a valence bias in credit assignment. If our truth bonus results were merely an artifact of positivity bias, this extended model should absorb that variance, producing a null truth bonus parameter. However, fitting this model to participant data still revealed a significant positive truth bonus, which again exceeds the range predicted by simulations of our Credibility CA model:

"3.7 Truth inference is still detected when controlling for valence bias

Given that participants frequently select bandits that are, on average, mostly rewarding, it is reasonable to assume that positive feedback is more likely to be objectively true than negative feedback. This raises a question if the "truth inference" effect we observed in participants might simply be an alternative description of a positivity bias in learning. To directly test this idea, we extended our Truth-CA model to explicitly account for a valence bias in credit assignment. This extended model features separate CA parameters for positive and negative feedback for each agent. When we fitted this new model to participant behavior, it still revealed a significant truth bonus in both the main study (Wilkoxon's signrank test: median = 0.09, $z(202)=2.12$, $p=0.034$; Fig. S27a) and the discovery study (median = 3.52, $z(102)=7.86$, $p<0.001$; Fig. S27c). Moreover, in the main study, this truth bonus remained significantly higher than what was predicted by all the alternative models, with the exception of the instructed-credibility bayesian model (Fig. S27b). In the discovery study, the truth bonus was significantly higher than what was predicted by all the alternative models (Fig. S27d)."

Together, these findings suggest that our truth inference results are not simply a re-description of a positivity bias.

Conversely, we acknowledge the reviewer's point that our positivity bias results could potentially stem from a more general truth inference mechanism. We believe that this possibility should be addressed in a future study where participants rate their belief that

received feedback is true (rather than a lie).We have extended our discussion to clarify this possibility and to include the suggested citation:

“Our findings show that individuals increase their credit assignment for feedback in proportion to the perceived probability that the feedback is true, even after controlling for source credibility and feedback valence. Strikingly, this learning bias was not predicted by any of our Bayesian or credit-assignment (CA) models. Notably, our evidence for this bias is based on a “oracle model” that incorporates the probability of feedback truthfulness from the experimenter's perspective, rather than the participant’s. This raises an important open question: how do individuals form beliefs about feedback truthfulness, and how do these beliefs influence credit assignment? Future research should address this by eliciting trial-by-trial beliefs about feedback truthfulness. Doing so would also allow for testing the intriguing possibility that an exaggerated positivity bias for non-credible sources reflects, to some extent, a truth-based discounting of negative feedback—i.e., participants may judge such feedback as less likely to be true. However, it is important to note that the positivity bias observed for fully credible sources (here and in other literature) cannot be attributed to a truth bias—unless participants were, against instructions, distrustful of that source.”

> *The authors get close to this in the discussion, but they characterize their results as differing from the predictions of rational models, the opposite of my intuition. They write:*
>
> *“Alternative "informational" (motivation-independent) accounts of positivity and confirmation bias predict a contrasting trend (i.e., reduced bias in low- and medium credibility conditions) because in these contexts it is more ambiguous whether feedback confirms one's choice or outcome expectations, as compared to a full-credibility condition.”*
>
> *I don't follow the reasoning here at all. It seems to me that the possibility for bias will increase with ambiguity (or perhaps will be maximal at intermediate levels). In the extreme case, when feedback is fully reliable, it is impossible to rationally discount it (illustrated in Figure 6A). The authors should clarify their argument or revise their conclusion here.*

We apologize for the lack of clarity in our previous explanation. We removed the sentence you cited (it was intended to make a different point which we now consider non-essential). Our current narration is consistent with the point you are making.

> *(4) Disinformation or less information?*
>
> *Zooming out, from a computational/functional perspective, the reliability of feedback is very similar to reward stochasticity (the difference is that reward stochasticity decreases the importance/value of learning in addition to its difficulty). I imagine that many of the effects reported here would be reproduced in that setting. To my surprise, I couldn't quickly find a study asking that precise question, but if the authors know of such work, it would be very useful to draw comparisons. To put a finer point on it, this study does not isolate which (if any) of these effects are specific to* disinformation*, rather than simply* less information*. I don't think the authors need to rigorously address this in the current study, but it would be a helpful discussion point.*

We thank the reviewer for highlighting the parallel (and difference) between feedback reliability and reward stochasticity. However, we have not found any comparable results in the literature. We also note that our discussion includes a paragraph addressing the locus of our effects making the point that more studies are necessary to determine whether our findings are due to disinformation per se or sources being less informative. While this

paragraph was included in the previous version it led us to infer our Discussion was too long and we therefore shortened it considerably:

"An important question arises as to the psychological locus of the biases we uncovered. Because we were interested in how individuals process disinformation—deliberately false or misleading information intended to deceive or manipulate—we framed the feedback agents in our study as deceptive, who would occasionally "lie" about the true choice outcome. However, statistically (though not necessarily psychologically), these agents are equivalent to agents who mix truth-telling with random "guessing" or "noise" where inaccuracies may arise from factors such as occasionally lacking access to true outcomes, simple laziness, or mistakes, rather than an intent to deceive. This raises the question of whether the biases we observed are driven by the perception of potential disinformation as deceitful per se or simply as deviating from the truth. Future studies could address this question by directly comparing learning from statistically equivalent sources framed as either lying or noisy. Unlike previous studies wherein participants had to infer source credibility from experience (30,37,72), we took an explicit-instruction approach, allowing us to precisely assess source-credibility impact on learning, without confounding it with errors in learning about the sources themselves. More broadly, our work connects with prior research on observational learning, which examined how individuals learn from the actions or advice of social partners (72–75). This body of work has demonstrated that individuals integrate learning from their private experiences with learning based on others' actions or advice—whether by inferring the value others attribute to different options or by mimicking their behavior (57,76). However, our task differs significantly from traditional observational learning. Firstly, our feedback agents interpret outcomes rather than demonstrating or recommending actions (30,37,72). Secondly, participants in our study lack private experiences unmediated by feedback sources. Finally, unlike most observational learning paradigms, we systematically address scenarios with deliberately misleading social partners. Future studies could bridge this by incorporating deceptive social partners into observational learning, offering a chance to develop unified models of how individuals integrate social information when credibility is paramount for decision-making."

> *(5) Over-reliance on analyzing model parameters*
>
> *Most of the results rely on interpreting model parameters, specifically, the "credit assignment" (CA) parameter. Exacerbating this, many key conclusions rest on a comparison of the CA parameters fit to human data vs. those fit to simulations from a Bayesian model. I've never seen anything like this, and the authors don't justify or even motivate this analysis choice. As a general rule, analyses of model parameters are less convincing than behavioral results because they inevitably depend on arbitrary modeling assumptions that cannot be fully supported. I imagine that most or even all of the results presented here would have behavioral analogues. The paper would benefit greatly from the inclusion of such results. It would also be helpful to provide a description of the model in the main text that makes it very clear what exactly the CA parameter is capturing (see next point).*

We thank the reviewer for this important suggestion which we address together with the following point.

> *(6) RL or regression?*
>
> *I was initially very confused by the "RL" model because it doesn't update based on the TD error. Consequently, the "Q values" can go beyond the range of possible reward (SI Figure 5). These values are therefore* not *Q values, which are defined as expectations of future reward ("action values"). Instead, they reflect choice propensities, which are sometimes notated h in the RL literature. This misuse of notation is unfortunately quite*

> *common in psychology, so I won't ask the authors to change the variable. However, they should clarify when introducing the model that the Q values are not action values in the technical sense. If there is precedent for this update rule, it should be cited.*
>
> *Although the change is subtle, it suggests a very different interpretation of the model.*
>
> *Specifically, I think the "RL model" is better understood as a sophisticated logistic regression, rather than a model of value learning. Ignoring the decay term, the CA term is simply the change in log odds of repeating the just-taken action in future trials (the change is negated for negative feedback). The PERS term is the same, but ignoring feedback. The decay captures that the effect of each trial on future choices diminishes with time. Importantly, however, we can re-parameterize the model such that the choice at each trial is a logistic regression where the independent variables are an exponentially decaying sum of feedback of each type (e.g., positive-cred50, positive-cred75, ... negative-cred100). The CA parameters are simply coefficients in this logistic regression.*
>
> *Critically, this is not meant to "deflate" the model. Instead, it clarifies that the CA parameter is actually not such an assumption-laden model estimate. It is really quite similar to a regression coefficient, something that is usually considered "model agnostic". It also recasts the non-standard "cross-fitting" approach as a very standard comparison of regression coefficients for model simulations vs. human data. Finally, using different CA parameters for true vs false feedback is no longer a strange and implausible model assumption; it's just another (perfectly valid) regression. This may be a personal thing, but after adopting this view, I found all the results much easier to understand.*

We thank the reviewer for their insightful and illuminating comments, particularly concerning the interpretation of our model parameters and the nature of our Credit assignment model. We believe your interpretation of the model is accurate and we now narrate it to readers in the hope that our modelling will become clearer and more intuitively. We also present to readers how these recasts our “cross-fitting” approach in the way you suggested (we return to this point below).

Broadly, while we agree that modelling results depend on underlying assumptions, we believe that “model-agnostic” approaches also have important limitations—especially in reinforcement learning (RL), where choices are shaped by histories of past events, which such approaches often fail to fully account for. As students of RL, we are frequently struck by how careful modelling demonstrates that seemingly meaningful “model-agnostic” patterns can emerge as artefacts of unaccounted-for variables. We also note that the term “model-agnostic” is difficult to define—after all, even regression models rely on assumptions, and some computational models make richer or more transparent assumptions than others. Ideally, we aim to support our findings using converging methods wherever possible.

We want to clarify that many of our reported findings indeed stem from straightforward behavioral analyses (e.g., simple regressions of choice-repetition), which do not rely on complex modeling assumptions. The two key results that primarily depend on the analysis of model parameters are our findings related to positivity bias and truth inference.

Regarding the positivity bias, identifying truly model-agnostic behavioral signatures, distinct from effects like choice-perseveration, has historically been a significant challenge in the literature. Classical research on this bias rests on the interpretation of model parameters (Lefebvre et al., 2017; Palminteri et al., 2017), or at least on the use of models to assess what an “unbiased learner” baseline should look like (Palminteri & Lebreton, 2022). Some researchers have suggested possible regressions incorporating history effects to detect positivity bias from choicerepetition behavior, but these regressions (as our model) rely on subtle assumptions about forgetting and history effects (Toyama et al., 2019). Specifically, in

our case, this issue is also demonstrated by analysis we conducted related to the previous point the reviewer made (about perseveration masquerading as positivity bias). We believe that dissociating clearly positivity bias from perseveration is an important challenge for the field going forward.

For our truth inference results, obtaining purely behavioral signatures is similarly challenging due to the intricate interdependencies (the reviewer has identified in previous points) between agent credibility, feedback valence, feedback truthfulness, and choice accuracy within our task design.

Finally, we agree with the reviewer that regression coefficients are often interpreted as a "modelagnostic" pattern. From this perspective even our findings regarding positivity and truth bias are not a case of over-reliance on complex model assumptions but are rather a way to expose deviations between empirical "sophisticated" regression coefficients and coefficients predicted from Bayesian models.

We have now described the main learning rule of our model in the main text to ensure that the meaning of the CA parameters is clearer for readers:

"Next, we formulated a family of non-Bayesian computational RL models. Importantly, these models can flexibly express non-Bayesian learning patterns and, as we show in following sections, can serve to identify learning biases deviating from an idealized Bayesian strategy. Here, an assumption is that during feedback, the choice propensity for the chosen bandit (which here is represented by a point estimate, "Q value", rather than a distribution) either increases or decreases (for positive or negative feedback, respectively) according to a magnitude quantified by the free "Credit-Assignment (CA)" model parameters (47):

$$Q(chosen) \leftarrow (1 - f_Q) * Q(chosen) + CA(agent, valence) * F$$

where F is the feedback received from the agents (coded as 1 for reward feedback and -1 for non-reward feedback), while fQ (∈[0,1]) is the free parameter representing the forgetting rate of the Q-value (Fig. 2a, bottom panel; Fig. S5b; Methods). The probability to choose a bandit (say A over B) in this family of models is a logistic function of the contrast choice-propensities between these two bandits. One interpretation of this model is as a "sophisticated" logistic regression, where the CA parameters take the role of "regression coefficients" corresponding to the change in log odds of repeating the just-taken action in future trials based on the feedback (+/- CA for positive or negative feedback, respectively; the model also includes gradual perseveration which allows for constant log-odd changes that are not affected by choice feedback; see "Methods: RL models") . The forgetting rate captures the extent to which the effect of each trial on future choices diminishes with time. The Q-values are thus exponentially decaying sums of logistic choice propensities based on the types of feedback a bandit received."

We also explain the implications of this perspective for our cross-fitting procedure:

"To further characterise deviations between behaviour and our Bayesian learning models, we used a "crossfitting" method. Treating CA parameters as data-features of interest (i.e., feedback dependent changes in choice propensity), our goal was to examine if and how empirical features differ from features extracted from simulations of our Bayesian learning models. Towards that goal, we simulated synthetic data based on Bayesian agents (using participants' best fitting parameters), but fitted these data using the CA-models, obtaining what we term "Bayesian-CA parameters" (Fig. 2d; Methods). A comparison of these BayesianCA parameters, with empirical-CA parameters obtained by fitting CA models to empirical data, allowed us to uncover patterns consistent with, or deviating from, ideal-Bayesian value-based inference. Under the sophisticated logistic-regression interpretation of the CA-model family the cross-fitting method comprises a comparison between empirical regression coefficients (i.e., empirical CA parameters) and regression coefficient based on

simulations of Bayesian models (Bayesian CA parameters). Using this approach, we found that both the instructed-credibility and free-credibility Bayesian models predicted increased BayesianCA parameters as a function of agent credibility (Fig. 3c; see SI 3.1.1.2 Tables S8 and S9). However, an in-depth comparison between Bayesian and empirical CA parameters revealed discrepancies from ideal Bayesian learning, which we describe in the following sections."

> ***Recommendations for the authors:***
>
> ***Reviewer #3 (Recommendations for the authors):***
>
> *(1) Keep terms consistent, e.g., follow-up vs. main; hallmark vs. traditional.*

We have now changed the text to keep terms consistent.

> *(2) CA model is like a learning rate; but it's based on the raw reward, not the TD error - this seems strange.*

We thank the reviewer for this comment. We understand that the use of a CA model instead of a TD error model may seem unusual at first glance. However, the CA model offers an important advantage: it more easily accommodates what we term "negative learning rates". This means that some participants may treat certain agents (especially the random one) as consistently deceitful, leading them to effectively increase/reduce choice tendencies following negative/positive feedback. A CA model handles this naturally by allowing negative CA parameters as a simple extension of positive ones. In contrast, adapting a TD error model to account for this is more complex. For instance, attempting to introduce a "negative learning rate" makes the RW model behave in a non-stable manner (e.g., Q values become <0 or >1). At the initial stages of our project, we explored different approaches to dealing with this issue and we found the CA model provides the best approach. For these reasons, we decided to proceed with our CA model.

Additionally, we used the CA model in previous studies (e.g., Moran, Dayan & Dolan (2021)) where we included (in SI) a detailed discussion of the similarities and difference between creditassignment and Rescorla-Wagner models

> *(3) Why was the follow-up study not pre-registered?*

We appreciate the reviewer's comment regarding preregistration, which we should have done. Unfortunately, this is now "water under the bridge" but going forward we hope to pre-register increasing parts of our work.

> *(4) Other work looking at reward stochasticity?*

As noted in point 4 of the main weaknesses, previous work on reward stochasticity primarily focused on explaining the increase/decrease in learning and its mechanistic bases under varying stochasticity levels. In our study, we uniquely characterize several specific learning biases that are modulated by source credibility, a topic not extensively explored within the existing reward stochasticity framework, as far as we know.

> *(5) Equation 1 is different from the one in the figure?*

The reviewer is completely correct. The figure provides a simplified visual representation, primarily focusing on the feedback-based update of the Q-value, and for simplicity, it omits the forgetting term present in the full Equation 1. To ensure complete clarity and prevent any misunderstanding, we have now incorporated a more detailed explanation of the model,

including the complete Equation 1 and its components, directly within the main text. This comprehensive description will ensure that readers are fully aware of how the model operates.

“Next, we formulated a family of non-Bayesian computational RL models. Importantly, these models can flexibly express non-Bayesian learning patterns and, as we show in following sections, can serve to identify learning biases deviating from an idealized Bayesian strategy. Here, an assumption is that during feedback, the choice propensity for the chosen bandit (which here is represented by a point estimate, “Q value“, rather than a distribution) either increases or decreases (for positive or negative feedback, respectively) according to a magnitude quantified by the free “Credit-Assignment (CA)” model parameters (47):

$$Q(chosen) \leftarrow (1 - f_Q) * Q(chosen) + CA(agent, valence) * F$$

where F is the feedback received from the agents (coded as 1 for reward feedback and -1 for non-reward feedback), while fQ (∈[0,1]) is the free parameter representing the forgetting rate of the Q-value (Fig. 2a, bottom panel; Fig. S5b; Methods).”

> *(6) Please describe/plot the distribution of all fitted parameters in the supplement. I would include the mean and SD in the main text (methods) as well.*

Following the reviewer’s suggestions, we have included in the Supplementary Document tables displaying the mean and SD of fitted parameters from participants for our main models of interest. We have also plotted the distributions of such parameters. Both for the main study:

> *(7) "A novel approach within the disinformation literature by exploiting a Reinforcement Learning (RL) experimental framework".*
>
> *The idea of applying RL to disinformation is not new. Please tone down novelty claims. It would be nice to cite/discuss some of this work as well.*
>
> *https://arxiv.org/abs/2106.05402?utm_source=chatgpt.com https://www.scirp.org/pdf/jbbs_2022110415273931.pdf https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4173312*

We thank the reviewer for pointing us towards relevant literature. We have now toned down the sentence in the introduction and cited the references provided:

“To address these questions, we adopt a novel approach within the disinformation literature by exploiting a Reinforcement Learning (RL) experimental framework (36). While RL has guided disinformation research in recent years (37–40), our approach is novel in using one of its most popular tasks: the “bandit task”.”

> *(8) Figure 3a - The figures should be in the order that they're referenced (3 is referenced before 2).*

We generally try to stick to this important rule but, in this case, we believe that our ordering serves better the narrative and hope the reviewer will excuse this small violation.

> *(9) "Additionally, we found a positive feedback-effect for the 3-star agent"*
>
> *What is the analysis here? To avoid confusion with the "positive feedback" effect, consider using "positive effect of feedback". The dash wasn't sufficient to avoid confusion in my case.*

We have now updated the terms in the text to avoid confusion.

> *(10) The discovery study revealed even stronger results supporting a conclusion that the credibility-CA model was superior to both Bayesian models for most subjects*
>
> *This is very subjective, but I'll just mention that my "cherry-picking" flag was raised by this sentence. Are you only mentioning cases where the discovery study was consistent with the main study? Upon a closer read, I think the answer is most likely "no", but you might consider adopting a more systematic (perhaps even explicit) policy on when and how you reference the discovery study to avoid creating this impression in a more casual reader.*

We thank the reviewer for this valuable suggestion. To prevent any impression of "cherry-picking", we have removed specific references to the discovery study from the main body of the text. Instead, all discussions regarding the convergence and divergence of results between the two studies are now in the dedicated section focusing on the discovery study:

“The discovery study (n=104) used a disinformation task structurally similar to that used in our main study, but with three notable differences: 1) it included 4 feedback agents, with credibilities of 50%, 70%, 85% and 100%, represented by 1, 2, 3, and 4 stars, respectively; 2) each experimental block consisted of a single bandit pair, presented over 16 trials (with 4 trials for each feedback agent); and 3) in certain blocks, unbeknownst to participants, the two bandits within a pair were equally rewarding (see SI section 1.1). Overall, this study's results supported similar conclusions as our main study (see SI section 1.2) with a few differences. We found convergent support for increased learning from more credible sources (SI 1.2.1), superior fit for the CA model over Bayesian models (SI 1.2.2) and increased learning from feedback inferred to be true (SI 1.2.6). Additionally, we found an inflation of positivity bias for low-credibility both when measured relative to the overall level of credit assignment (as in our main study), or in absolute terms (unlike in our main study) (Fig. S3; SI 1.2.5). Moreover, choice-perseveration could not predict an amplification of positivity bias for low-credibility sources (see SI 3.6.2). However, we found no evidence for learning based on 50%-credibility feedback when examining either the feedback effect on choice repetition or CA in the credibility-CA model (SI 1.2.3).”

> *(11) An in-depth comparison between Bayesian and empirical CA parameters revealed discrepancies from normative Bayesian learning.*
>
> *Consider saying where this in-depth comparison can be found (based on my reading, I think you're referring to the next section?*

We have now modified the sentence for better clarity:

“However, an in-depth comparison between Bayesian and empirical CA parameters revealed discrepancies from ideal Bayesian learning, which we describe in the following sections.”

> *(12) "which essentially provides feedback" Perhaps you meant "random feedback"?*

We have modified the text as suggested by the reviewer.

<(13) Essentially random

> *Why "essentially"? Isn't it just literally random?*

We have modified the text as suggested by the reviewer.

> *(14) Both Bayesian models predicted an attenuated credit-assignment for the 3-star agent*
>
> *Attenuated relative to what? I wouldn't use this word if you mean weaker than what we see in the human data. Instead, I would say people show an* exaggerated *credit-assignment, since Bayes is the normative baseline.*

We changed the text according to the reviewer's suggestion:

"A comparison of empirical and Bayesian credit-assignment parameters revealed a further deviation from ideal Bayesian learning: participants showed an exaggerated credit-assignment for the 3-star agent compared with Bayesian models."

> *(15) "there was no difference between 2-star and 3-star agent contexts (b=0.051, F(1,2419)=0.39, p=0.53)"*
>
> *You cannot confirm the null hypothesis! Instead, you can write "The difference between 2-star and 3-star agent contexts was not significant". Although even with this language, you should be careful that your conclusions don't rest on the lack of a difference (the next sentence is somewhat ambiguous on this point).*
>
> *Additionally, the reported b coefs do not match the figure, which if anything, suggests a larger drop from 0.75 (2-star) to 1 (3-star). Is this a mixed vs fixed effects thing? It would be helpful to provide an explanation here.*

We thank the reviewer for this question. When we previously submitted our manuscript, we thought that finding enhanced credit-assignment for fully credible feedback following potential disinformation from a DIFFERENT context would constitute a striking demonstration of our "contrast effect". However, upon reexamining this finding we found out we had a coding error (affecting how trials were filtered). We have now rerun and corrected this analysis. We have assessed the contrast effect for both "same-context" trials (where the contextual trial featured the same bandit pair as the learning trial) and "different-context" trials (where the contextual trial featured a different bandit pair). Our re-analysis reveals a selective significant contrast effect in the same-context condition, but no significant effect in the different-context condition. We have updated the main text to reflect these corrected findings and provide a clearer explanation of the analysis:

"A comparison of empirical and Bayesian credit-assignment parameters revealed a further deviation from ideal Bayesian learning: participants showed an exaggerated credit-assignment for the 3-star agent compared with Bayesian models [Wilcoxon signed-rank test, instructed-credibility Bayesian model (median difference=0.74, z=11.14); free-credibility Bayesian model (median difference=0.62, z=10.71), all p's<0.001] (Fig. 3a). One explanation for enhanced learning for the 3-star agents is a contrast effect, whereby credible information looms larger against a backdrop of non-credible information. To test this hypothesis, we examined whether the impact of feedback from the 3-star agent is modulated by the credibility of the agent in the trial immediately preceding it. More specifically, we reasoned that the impact of a 3-star agent would be amplified by a "low credibility context" (i.e., when it is preceded by a low credibility trial). In a binomial mixed effects model, we regressed choice-repetition on feedback valence from the last trial featuring the same bandit pair (i.e., the learning trial) and the feedback agent on the trial immediately preceding that last trial (i.e., the contextual credibility; see Methods for model-specification). This analysis included only learning trials featuring the 3-star agent, and context trials featuring the same bandit pair as the learning trial (Fig. 4a). We found that feedback valence interacted with contextual credibility ($F(2,2086)=11.47$, $p<0.001$) such that the feedback-effect (from the 3-star agent) decreased as a function of the preceding context-credibility (3-star context vs. 2-star context:

b= -0.29, F(1,2086)=4.06, p=0.044; 2star context vs. 1-star context: b=-0.41, t(2086)=-2.94, p=0.003; and 3-star context vs. 1-star context: b=0.69, t(2086)=-4.74, p<0.001) (Fig. 4b). This contrast effect was not predicted by simulations of our main models of interest (Fig. 4c). No effect was found when focussing on contextual trials featuring a bandit pair different than the one in the learning trial (see SI 3.5). Thus, these results support an interpretation that credible feedback exerts a greater impact on participants' learning when it follows non-credible feedback, in the same learning context."

We have modified the discussion accordingly as well:

"A striking finding in our study was that for a fully credible feedback agent, credit assignment was exaggerated (i.e., higher than predicted by our Bayesian models). Furthermore, the effect of fully credible feedback on choice was further boosted when it was preceded by a low-credibility context related to current learning. We interpret this in terms of a "contrast effect", whereby veridical information looms larger against a backdrop of disinformation (21). One upshot is that exaggerated learning might entail a risk of jumping to premature conclusions based on limited credible evidence (e.g., a strong conclusion that a vaccine produces significant side-effect risks based on weak credible information, following non-credible information about the same vaccine). An intriguing possibility, that could be tested in future studies, is that participants strategically amplify the extent of learning from credible feedback to dilute the impact of learning from noncredible feedback. For example, a person scrolling through a social media feed, encountering copious amounts of disinformation, might amplify the weight they assign to credible feedback in order to dilute effects of 'fake news'. Ironically, these results also suggest that public campaigns might be more effective when embedding their messages in low-credibility contexts, which may boost their impact."

And we have included some additional analyses in the SI document:

"3.5 Contrast effects for contexts featuring a different bandit Given that we observed a contrast effect when both the learning and the immediately preceding "context trial" involved the same pair of bandits, we next investigated whether this effect persisted when the context trial featured a different bandit pair – a situation where the context would be irrelevant to the current learning. Again, we used in a binomial mixed effects model, regressing choice-repetition on feedback valence in the learning trial and the feedback agent in the context trial. This analysis included only learning trials featuring the 3-star agent, and context trials featuring a different bandit pair than the learning trial (Fig. S22a). We found no significant evidence of an interaction between feedback valence and contextual credibility (F(2,2364)=0.21, p=0.81) (Fig. S22b). This null result was consistent with the range of outcomes predicted by our main computational models (Fig. S22c)."

We aimed to formally compare the influence of two types of contextual trials: those featuring the same bandit pair as the learning trial versus those featuring a different pair. To achieve this, we extended our mixedeffects model by incorporating a new predictor variable, "CONTEXT_TYPE" which coded whether the contextual trial involved the same bandit pair (coded as -0.5) or a different bandit pair (+0.5) compared to the learning trial. The Wilkinson notation for this expanded mixed-effects model is:

$$REPEAT \sim CONTEXT\_TYPE * FEEDBACK * (C\ CONTEXT_{2\text{-star}} + CONTEXT_{3\text{-star}}) + BETTER + (1 \mid participant)$$

This expanded model revealed a significant three-way interaction between feedback valence, contextual credibility, and context type (F(2,4451) = 7.71, p<0.001). Interpreting this interaction, we found a 2-way interaction between context-source and feedback valence when the context was the same (F(2,4451) = 12.03, p<0.001), but not when context was different (F(2,4451) = 0.23, p = 0.79). Further interpreting the double feedback-valence *

context-source interaction (for the same context) we obtained the same conclusions as reported in the main text."

> *(16) "Strikingly, model-simulations (Methods) showed this pattern is not predicted by any of our other models"*
>
> *Why doesn't the Bayesian model predict this?*

Thanks for the comment. Overall, Bayesian models do predict a slight truth inference effect (see Figure 6d). However, these effects are not as strong as the ones observed in participants, suggesting that our results go beyond what would be predicted by a Bayesian model.

Conceptually, it's important to note that the Bayesian model can infer (after controlling for source credibility and feedback valence) whether feedback is truthful based solely on prior beliefs about the chosen bandit. Using this inferred truth to amplify the weight of truthful feedback would effectively amount to "bootstrapping on one's own beliefs." This is most clearly illustrated with the 50% agent: if one believes that a chosen bandit yields rewards 70% of the time, then positive feedback is more likely to be truthful than negative feedback. However, a Bayesian observer would also recognize that, given the agent's overall unreliability, such feedback should be ignored regardless.

> *(17) "A striking finding in our study was that for a fully credible feedback agent, credit assignment was exaggerated (i.e., higher than predicted by a Bayesian strategy)".*
>
> *"Since we did not find any significant interactions between BETTER and the other regressors, we decided to omit it from the model formulation".*
>
> *Was this decision made after seeing the data? If so, please report the original analysis as well.*

We have included the BETTER regressor again, and we have re-run the analyses. We now report the results of such regression. We have also changed the methods section accordingly:

"We used a different mixed-effects binomial regression model to test whether value learning from the 3-star agent was modulated by contextual credibility. We focused this analysis on instances where the previous trial with the same bandit pair featured the 3-star agent. We regressed the variable REPEAT, which indicated whether the current trial repeated the choice from the previous trial featuring the same bandit-pair (repeated choice=1, non-repeated choice=0). We included the following regressors: FEEDBACK coding the valence of feedback in the previous trial with the same bandit pair (positive=0.5, negative=-0.5), CONTEXT2-star indicating whether the trial immediately preceding the previous trial with the same bandit pair (context trial) featured the 2-star agent (feedback from 2-star agent=1, otherwise=0), and CONTEXT3star indicating whether the trial immediately preceding the previous trial with the same bandit pair featured the 3-star agent. We also included a regressor (BETTER) coding whether the bandit chosen in the learning trial was the better -mostly rewarding- or the worse -mostly unrewarding- bandit within the pair. We included in this analysis only current trials where the context trial featured a different bandit pair. The model in Wilkinson's notation was:

$$REPEAT \sim FEEDBACK * (CONTEXT_{2\text{-star}} + CONTEXT_{3\text{-star}}) + BETTER + (1 \mid participant) \quad (13)$$

In figure 4c, we independently calculate the repeat probability difference for the better (mostly rewarding) and worse (mostly non-rewarding) bandits and averaged across them. This calculation was done at the participants level, and finally averaged across participants."

https://doi.org/10.7554/eLife.106073.2.sa4