Biased Misinformation Distorts Beliefs

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

Affiliations:

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

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

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

***Corresponding authors**. Email: rani.moran@gmail.com, juanvidalpe@gmail.com

ABSTRACT

We often form beliefs about events we cannot directly observe by relying on information provided by others. However, these sources may be biased, potentially distorting our beliefs and behavior. Here, we examine whether individuals can detect biased sources and mitigate their influence. We studied a large cohort of participants who completed a decision-making task involving choices between lotteries. Outcome feedback was provided by sources that were either unbiased or biased in a favorable or unfavorable direction. Participants initially learned about source biases, and later, used this knowledge to interpret biased feedback and adjust belief updating. Using a reinforcement learning framework, we show that participants successfully distinguished between favorable, unfavorable, and unbiased sources and adjusted for biased feedback. However, these corrections were incomplete, allowing residual biases to continue shaping beliefs and decisions. Moreover, following exposure to biased sources, participants systematically misperceived unbiased sources as biased. Strikingly, participants prioritized learning about source biases over value learning, resulting in reduced task performance. Our findings highlight the challenge of forming accurate beliefs in biased environments and offer insights for countering misinformation.

INTRODUCTION

Accurate information enables people to adapt effectively to their environment. Much information we receive is mediated by others, rendering it prone to inaccuracies. Biased information, involving systematic deviations from objective truth, plays a major role in the adherence to, and spread of, misinformation (1). Biases can be intentional or unintentional, such as selective fact presentation (e.g., political media cherry-picking evidence), subjective interpretations (e.g., exaggerating the probability of a virus spreading), or outright use of manipulative tactics (e.g., sellers inflating product quality). Uncorrected, biased misinformation can distort an individual's beliefs, resulting in political radicalization (2,3), poor health decisions (4–6), and adherence to conspiracy theories (7,8).

Misinformation has emerged as a problematic issue in today's information age, where biased media sources are pervasive (9–11), exacerbating societal polarization (12) and leading to entrenched false beliefs (13–15). Moreover, training AI systems on corpora of biased human data embeds those biases into the very operations of these systems, perpetuating and amplifying human biases (16,17). Large language models are also known to show stereotype biases (18), raising a question as to whether, and how, accurate belief formation (i.e., learning) can occur when so much information in our social environment is biased. Indeed, the ubiquitous prevalence of mis- and disinformation has sparked concern, as free democracies flourish when an informed citizenry provides the bedrock for political decision-making (19).

Research on individual learning has mostly examined how belief-updating is impacted by information precision, typically manipulated by adding unbiased noise (e.g., normally distributed) to the information individuals receive (20–25). Here, a key finding is that individuals decrease their learning rates as a function of imprecision (i.e., input noise). However, information corruption due to systematic-bias, as opposed to unbiased noise, presents a fundamentally different learning-challenge (26–31). Unlike unbiased noise, which introduces random variation, bias systematically skews data in predictable ways. Rather than reducing learning, in principle, individuals should be able to model, identify, and correct ("debias") such distortions. Yet, despite the importance of this distinction there is a dearth of research on how biased information shapes learning, and bias is often treated as a proxy for general credibility (14,32).

We conjectured that, through repeated interactions with an information source, individuals can compare a source's claims with either ground truth or personal experience, and gradually form an effective knowledge representation of this bias (33). Such a representation could then enable a debiased interpretation when the ground truth is not readily available. For example, knowing that your academic colleague tends to exaggerate paper-acceptance probabilities by 20%, you might adjust her prediction of a 90% likelihood of manuscript acceptance to a more realistic 70%. In simple terms, if individuals can perfectly estimate and correct for information-bias, they should be able to learn equally well when informed by (equally noisy) biased or unbiased sources. However, it may be challenging to apply these corrections accurately, as cognitive effort is required to correct biases (34). Furthermore, phenomena such as truth bias (35) or anchoring (36) may hinder a full correction (33). Additionally, motivated cognition theory suggests individuals are less inclined to correct favorably biased information, that portrays them in a positive light or confers positive

consequences, as compared to unfavorably biased information (37,38). Taking account of these complexities, our study asks whether, and to what extent, people accurately estimate the biases of information sources and use this knowledge to interpret information in the absence of an accessible ground truth.

We applied a Reinforcement Learning (RL) framework to ascertain how individuals learn from reward feedback in contexts involving misinformation (20,39). Responding to a recent call to enhance research methods for studying the impact of misinformation (40), this framework can quantify how individuals dynamically update their beliefs through repeated interactions with biased information sources, as well as how cognitive biases influence learning and decision-making (37,41,42). Moreover, RL paradigms share key features with misinformation-rich social media environments, including operating across short timescales where decisions, such as reposting a tweet (social media) or selecting a lottery (RL task), are enacted within seconds and shaped by reinforcers, such as monetary rewards in RL and likes or shares on social media (43–45).

We developed a novel “biased” variant of a popular multi-armed bandit RL task, mimicking scenarios where people form beliefs based on information from potentially biased sources. In the task participants chose between lotteries (bandits) of varying values based on choice reward-feedback from a range of potentially biased sources. This involved 'favorable' sources that exaggerated choice reward-outcomes positively, 'unfavorable' sources that skewed outcomes downward, and 'neutral' sources that provided unbiased feedback. Participants were not explicitly informed of these biases but could infer them during an initial phase of our task, where we provided true outcomes alongside biased feedback. In a subsequent phase, where true outcomes were withheld, participants could potentially rely on acquired knowledge of source-biases to correct (debias) biased feedback. Crucially, the biases were simple additive shifts, allowing for a straightforward debiasing strategy involving adding a constant correction to the feedback.

We show that, by comparing source-feedback with ground-truth, individuals acquire knowledge of information-source biases, allowing them to classify these biases and adjust source-feedback accordingly. In effect, they amplify input from unfavorable sources and attenuate input from favorable ones. However, exposure to biased misinformation also fostered biased beliefs with respect to both bandit-values and source-bias. First, despite its simplicity, a debiasing strategy remained incomplete, allowing residual source-biases to shape beliefs about bandit-values and influence decisions (e.g., bandits were perceived as more valuable and chosen more following interactions with a favorably biased agent). Second, the perception of neutral information sources manifested a “contrast bias” whereby, following interactions with a biased source (e.g., favorable), a neutral source was perceived as biased in the opposite direction (e.g., unfavorable). Finally, prioritizing learning about source-biases depleted cognitive resources, compromising reward acquisition in our task. Our findings highlight surprising complexities when learning from biased misinformation and reveal core mechanisms underlying belief formation in such environments.

RESULTS

We studied 200 participants who completed an online, multi-armed reinforcement learning task. In the task, participants were provided with a cover story (full detail in Methods), where they were told they would manage an art gallery which owned multiple copies of various paintings. Each round involved selecting between two offered paintings, as in a bandit task. A copy of the chosen painting was then sent to an auction house to be sold at a variable price based on its quality. For each painting the selling price of copies varied according to a normal distribution, with a mean randomly assigned (independently across paintings) to be either $17, $20, $23, $26, or $29, with a standard deviation of $1.

Participants were informed that they would collaborate with different, potentially biased, art agencies when making their selections. Upon selection of a painting, one of the agencies would provide an estimated selling price for its copy (“source feedback”), which could differ from the actual auction-house selling price (“true outcome”). Importantly source-biases were framed in a motivation-free manner: participants were explicitly told that the agencies vary in prices of the art they previously dealt with, and these past experiences might bias them to overestimate, underestimate, or offer relatively accurate estimates of selling prices.

Unbeknownst to participants, each agency belonged to one of 3 “bias types”: favorable agencies provided estimates that were on average $3 above the true selling price, unfavorable agencies $3 below, and neutral agencies around the actual price. Estimates from all agencies had a standard deviation of $3 (i.e., $agency\ feedback\ =\ true\ outcome\ +\ N(\mu, 3^2)$, with $\mu = -3,0,+3$ for unfavorable, neutral and favorable agencies respectively) (Figure 1a).

Participants completed six "superblocks" (Figure 1b), each containing 3 blocks of 24 trials with each block featuring copies of 4 new paintings. Each superblock featured two new agencies who always differed in their bias tendency (e.g., if one was favorable the other was either neutral or unfavorable). The three possible pairings of agencies bias-types were counterbalanced across superblocks, with each pairing being featured in two superblocks. In the initial block (i.e., phase 1 of the superblock), participants saw, on each trial, the painting's true selling price followed by the agency's feedback (Figure 1c). Participants were not instructed about the agencies’ bias but instead could estimate these based on a comparison of true outcomes with the agent’s feedback, allowing them to acquire knowledge of each agency’s bias. Following this block, participants classified each of the two agency’s biases as "above," "below," or "around" the selling price, and indicated their confidence in each classification.

Participants then played two additional blocks (i.e., phase 2), each featuring different sets of paintings, in which only the agency's feedback was shown, but where the true selling price remained hidden (Figure 1d). Participants were instructed that although the true selling prices were unavailable, their bonus payment still depended on these true latent outcomes (rather than agents’ estimates), thereby incentivizing learning of the true selling price of paintings based on biased feedback from the agencies. Afterwards, participants provided a second classification of

each agency's bias (again accompanied by confidence ratings) before beginning a new superblock featuring two new agencies.

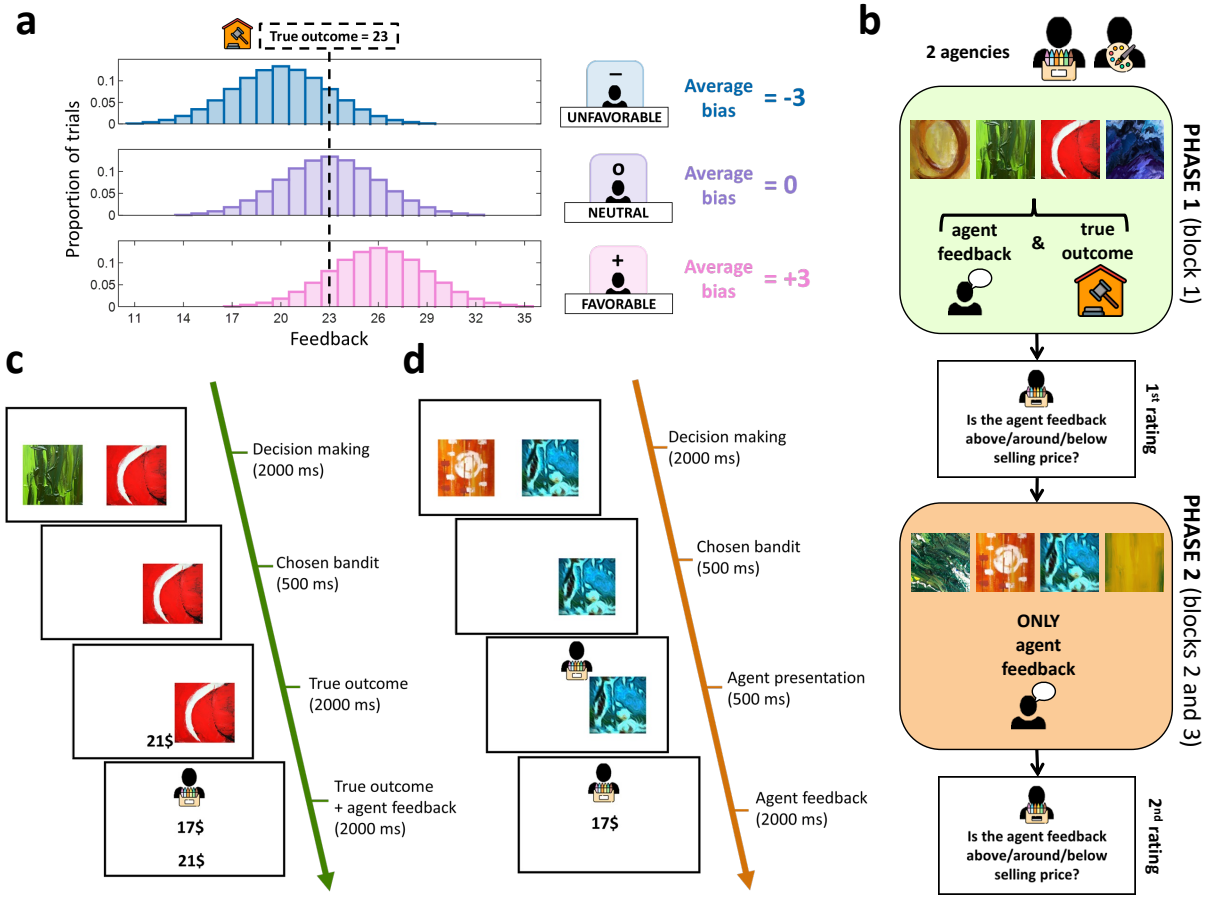

Figure 1: Task design. (a) Distribution of feedback from each agency illustrated for a true auction-house selling price of 23$. Agencies provided systematically biased feedback, relative to the true outcomes of a choice: unfavorable agents gave feedback consistently below the true value, neutral agents centered around the true value and favorable agents above the true value. Agency biases were sampled from Gaussian distributions with distinct means but the same standard deviation. Mean biases were set to -3 for unfavorable agents, 0 for neutral agents, and +3 for favorable agents. **(b)** Superblock structure. Each superblock involved participants interactions with two agencies across two phases. During Phase 1, they observed both the true outcomes of their choices, followed by the corresponding (biased) agency-feedback, allowing them to learn about agency biases. At the end of Phase 1, participants classified each agency's bias as well as rated their confidence in these classifications. During Phase 2, participants made choices but received only agency feedback, without access to true outcomes (note that blocks 2 and 3 included different painting quartets). At the end of Phase 2, participants classified agencies and reported confidence again. **(c)** Phase 1 trial structure. On each trial in Phase 1, participants first made a choice between a pair of pictures. The true outcome was displayed for 2000 ms, followed by the agency's biased feedback for another 2000 ms. **(d)** Phase 2 trial structure. On each trial in Phase 2, participants made a choice and saw only the biased feedback for 2000 ms, with the true outcome hidden. Unfavorable, neutral, and favorable agencies are denoted by the symbols "-" (blue), "o" (violet), and "+" (pink), respectively.

People imperfectly ‘debias’ biased feedback

We first assessed how biased feedback was used for value learning during the second phase of superblocks, where the true selling price of paintings was unavailable. Ideally, participants should adjust estimates from biased information-sources to counteract their respective biases - subtracting 3$ from favorable feedback and adding 3$ to unfavorable feedback, enabling consistent value learning across sources.

To examine whether individuals debiased feedback perfectly, we relied on a hallmark of RL value learning involving a tendency to repeat choices as a function of choice-outcome. If individuals perfectly debias feedback, then (debiased) feedback from all sources will be equally distributed, resulting in an equal probability of repeating on the next (n+1) trial the current (n) trial’s choice across information-sources (i.e., the current trial’s source-bias type becomes irrelevant). By contrast, differences in choice repetition following feedback from different sources implicates imperfect debiasing. We tested this in a binomial logistic mixed-effects model, regressing choice repetition (i.e., whether participants repeated their choice of a painting when it is offered on the next trial n+1, coded as 1 for repeat and 0 for non-repeat) on source-bias type (unfavorable, neutral, or favorable) at the current trial n (see Methods for model specifications). We found that choice repetition differed across source-types ($F(2,1794) = 42.07$, $p < .001$), with lower repetition rates following feedback from unfavorable sources ($b = -0.18$, $t(1794) = -5.08$, $p < .001$), and higher following feedback from favorable sources ($b = 0.19$, $t(1794) = 5.07$, $p < .001$), when compared to neutral sources (Figure 2a). These findings indicate biased feedback was not perfectly adjusted for, allowing feedback sources’ biases to distort participants’ beliefs (e.g., perceiving a painting as more valuable following favorably biased as compared to unbiased feedback).

To investigate this further, we developed a family of computational models in which the latent value of each painting/bandit (represented as a Q-value) was updated based on source feedback. During each trial's feedback stage (on Phase 2), the source feedback was corrected by subtracting a source-specific debias parameter:

$$debiased\ feedback = feedback - debias(source) \quad \textbf{(1)}$$

where *debias* is a free parameter representing the value used to debias the feedback from each source. The debiased feedback was then used to update the Q-value of the selected painting according to a Rescorla-Wagner learning rule:

$$Q \leftarrow Q + \alpha * (debiased\ feedback - Q) \quad \textbf{(2)}$$

where α ($\in$[0,1]) is the free parameter representing the learning rate used to update the Q-value of the chosen painting (see Methods for a detailed model description).

We formulated three different model-versions. All versions allowed for a free baseline debias parameter corresponding to the neutral source, accounting for a possibility that neutral feedback

may be perceived as biased. Importantly, the three variants differed on how the debiasing magnitude was influenced by source-favorability. In the "consistent debias" model, the debias parameter for the unfavorable and favorable sources were fixed to its objective value relative to the debias for the neutral source ($\mathrm{debias(unfavorable)} = \mathrm{debias(neutral)} - 3$ and $\mathrm{debias(favorable)} = \mathrm{debias(neutral)} + 3$). In the "constant" model, all debias parameters were fixed to the baseline debias ($\mathrm{debias(unfavorable)} = \mathrm{debias(neutral)} = \mathrm{debias(favorable)}$), meaning that bias-favorability had no influence on debiasing. Lastly, in our "free debias" model, the debias parameters for all three feedback sources were treated as free parameters, allowing for a possibility that the magnitude of corrections could be freely modulated by the type of source-bias (we also examined in Figure S3 “0- debias parameters” variants of models where the baseline, neutral-source, bias was fixed to 0). We tested the predictive accuracy of our different models by first fitting each version to participants' behavior. Next, we simulated synthetic datasets based on the maximum likelihood (ML) parameters from participants, and then repeated the above choice repetition analysis on these simulated data (see Methods).

We found, unlike for the empirical data, the “consistent debias” model failed to predict choice-repetition sensitivity to source-favorability ($F(2,8987) = 0.14$, $p = 0.87$), supporting a conclusion that participants do not debias biased feedback with perfect consistency (Figure 2b, left panel). In contrast, both the “free debias” model ($F(2,8988) = 52.02$, $p < .001$; effect of unfavorable source: $b = -0.14$, $t(8988) = -5.51$, $p < .001$; effect of favorable source: $b = 0.14$, $t(8988) = 5.51$, $p < .001$; Figure 2b middle panel) and the “constant” model ($F(2,8992) = 195.0$, $p < .001$; effect of unfavorable source: $b = -0.24$, $t(8992) = -13.46$, $p < .001$; effect of favorable source: $b = 0.24$, $t(8992) = 12.94$, $p < .001$; Figure 2b, right panel) predicted the observed source effects on choice-repetition. This raises a question as to whether participants are simply insensitive to the type of source-bias in their debiasing, or instead, they are sensitive to source-bias type but under-debias (e.g., adding just 2$, instead of the optimal 3$, to unfavorable feedback, relative to the baseline).

Individuals undercorrect for biased feedback

We reasoned that if participants’ debiasing was constant across sources (i.e., they correct all sources in the same way), then (biased) choice-feedback alone, but not source-bias type, should affect choice-repetition. Conversely, source-type effects (after controlling for feedback) would implicate differential debiasing. To test this, we again used a binomial logistic mixed-effects model, this time regressing choice repetition at the next trial (n+1), on the type of source-bias and the feedback from the source, both from the current trial n (see Methods for model specifications). Results showed a positive effect of feedback on choice repetition ($b = 0.11$, $t(26184) = 20.88$, $p < .001$), indicating participants were learning based on feedback from the sources. Critically, we found a source type effect on choice repetition ($F(2,26184) = 12.69$, $p < .001$), with unfavorable sources showing a positive effect on repetition ($b = 0.17$, $t(26184) = 3.85$, $p < .001$), and favorable sources showing a negative effect on repetition ($b = -0.08$, $t(26184) = -2.01$, $p = 0.045$) - both compared to neutral sources (Figure 2c, left panel). This suggests that nominal feedback was adjusted upwards (relative to baseline) when it came from an unfavorable source and adjusted downwards when it came from a favorable source.

Model-based simulations also supported a conclusion that participants were sensitive to differences in bias across feedback sources. Indeed, while the “constant” model predicted a

positive main effect for feedback on repetition ($b = 0.088$, $t(130400) = 26.35$, $p < .001$), it failed to predict an effect of source ($F(2,130405) = 0.76$, $p = 0.47$; Figure 2c, top-right panel). In contrast, and similar to our empirical data, the "free debias" model predicted main effects for both feedback ($b = 0.096$, $t(130780) = 26.15$, $p < .001$), and source-bias type ($F(2,130780) = 33.41$, $p < .001$; effect of unfavorable sources: $b = 0.13$, $t(130780) = 4.32$, $p < .001$; effect of favorable sources: $b = -0.14$, $t(130780) = -5.36$, $p < .001$) (Figure 2c, bottom-right panel).

Our findings highlight that the free-debias model is superior in accounting for data compared to both its submodels. Convergent evidence comes from a model comparison based on a bootstrap generalized likelihood ratio test (see Methods), which at the group-level rejects both the consistent and constant debias submodels variants in favor of the free-debias supermodel (both $p< .001$). This implies that while participants were sensitive to source-bias type when debiasing feedback, their debiasing patterns were not perfectly consistent across sources.

Next, to characterize the extent to which individuals (imperfectly) debias biased feedback, we examined the maximum likelihood (ML) debias parameters from our "free debias" model (Figure 2d; we ensured parameters of interest have good recoverability properties see Figure S1). A mixed-effects model, where we regressed these parameters on their associated source-type, revealed the debias for unfavorable sources was significantly negative ($b = -0.95$\$, $t(597) = -2.14$, $p = 0.033$), while the debias for favorable sources was significantly positive ($b = 1.90$\$, $t(597) = 4.28$, $p < .001$), both relative to the debias for neutral sources (see Methods). Hence, while participants debias feedback in the “correct direction”, the debias parameters reveal under-debiasing, with the unfavorable-source debias being above -3\$ ($F(1,597) = 21.26$, $p < .001$), and the favorable-source debias being below 3\$ ($F(1,597) = 6.08$, $p = 0.014$), both compared to neutral. Thus, it appears that participants are sensitive to feedback biases but their corrections for these are partial rather than full (see Figure S2a-b for similar conclusions when controlling for participants beliefs about the bias of each source). We found no evidence for difference in the extent of downward adjustment of favorable feedback and upward adjustments of unfavorable feedback ($F(1,597) = 1.53$, $p = 0.22$).

Absence of bias effects on learning rate

Prior research indicates that increases in unbiased feedback noise (i.e., precision) decreases learning rates (20–25). In our task, all information sources provided equally noisy information (Fig. 1a), though it might be that participants consider biased feedback as noisier and filter it out by a learning rate adjustment. To test this, we extended our free-debias model to allow learning rates to vary across source-bias types (see Methods). In a linear mixed-effects model, that regressed ML learning rate parameters on source type, we found no significant differences in learning rate between sources ($F(597,2) = 0.57$, $p = 0.57$) (Figure 2e; see Figure S2c for debias parameters in this extended model). This suggests that rather than simply reducing a reliance on biased information, participants actively attempted to correct for bias—albeit with partial success—in order to infer less biased estimates of bandit values.

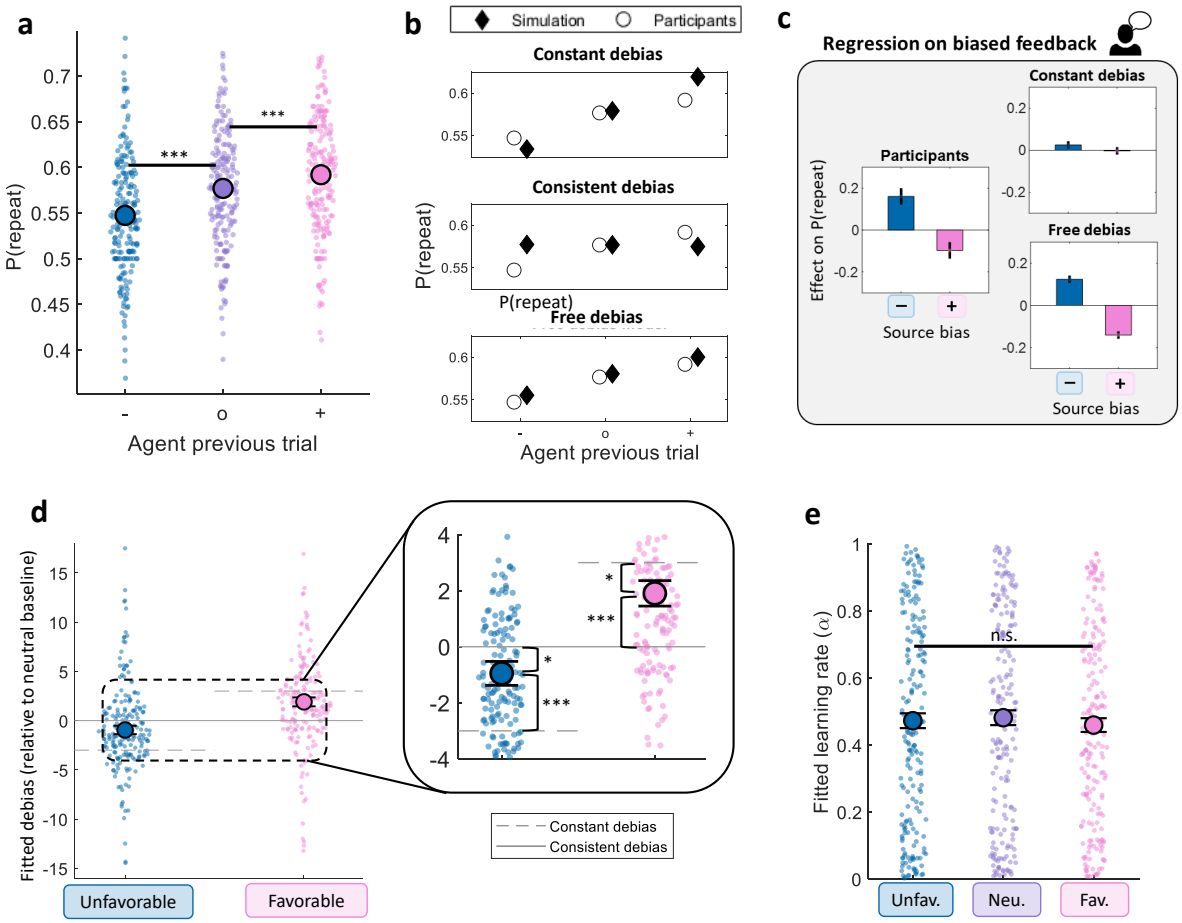

Figure 2. Characterization of source debiasing. (a) Probability of repeating a choice (when offered on two consecutive trials) as a function of the source providing feedback in the previous trial. Choice repetition was higher following favorable feedback, and lower following unfavorable feedback, relative to neutral feedback. **(b)** Choice repetition effects in simulated data, generated using maximum-likelihood (ML) parameters from participants. The "constant" model (top) assumes that bias-favorability has no influence on debiasing (i.e., debias for favorable/unfavorable sources is equal to the neutral baseline) and shows higher repetition following favorable feedback and lower repetition following unfavorable feedback. The "consistent debias" model (middle) debias parameter for the unfavorable and favorable sources were fixed to its objective value relative to the neutral source (*debias(neutral)-3* and *debias(neutral)+3* for favorable and unfavorable sources respectively), predicting equal choice repetition across sources. The "free debias" model (bottom) treats debiasing as free parameters and most accurately reproduces participants' choice repetition patterns. **(c)** Mixed effects modelling of choice repetition based on biased feedback and the type of source. Source-type effects are presented. Participants showed a positive/negative effect for unfavorable/favorable sources (left), indicating increased/decreased choice repetition for the same feedback when provided by unfavorable/favorable sources. This effect was not captured by the "constant" model (top right). **(d)** ML estimates of debias parameters for favorable and unfavorable feedback relative to neutral feedback. Group level parameters were between 0 (constant) and their ideal values (-3 for unfavorable and +3 for favorable), indicating that participants adjusted for feedback bias in the correct direction but undercorrected. **(e)** ML estimates of learning rate parameters for the different sources. The learning rate did not significantly differ across sources. Small dots represent individual participants/simulations, while large circles show group means. Error bars indicate standard errors of the

mean (SEM). Unfavorable, neutral, and favorable sources are denoted by the symbols “-” (blue), “o” (violet), and “+” (pink), respectively. (*) $p < .05$, (**) $p < .01$,(***) $p < .001$, (n.s.) $p>.05$.

Beliefs about source bias are linked to debias during learning

A subset of participants (n=188) were asked to classify the sources according to whether they tended to overestimate, underestimate or provide roughly accurate price-estimates, following phase 1 and phase 2 of each superblock (Figure 1b). Classification accuracy rates (Figure 3a) exceeded chance-level (0.333) for all sources (unfavorable: mean = 0.71, $t(187) = 18.59$; neutral: mean = 0.62, $t(187) = 15.63$; favorable: mean = 0.69, $t(187) = 16.18$; all p’s $< .001$). Using a binomial logistic mixed-effects model, we regressed source-classification accuracy rates on the source-bias type (favorable, neutral, unfavorable) and time of report (after phase 1 = -0.5, after phase 2 = 0.5) (see Methods for model specifications). This regression revealed a significant main effect for source-bias type on classification accuracy rates ($F(2,1122) = 10.60$, $p < .001$), with biased sources being classified more accurately than neutral sources (unfavorable source effect = 0.72, $t(1122) = 4.61$, $p < .001$; favorable source effect = 0.57, $t(1122) = 3.43$, $p < .001$) but with no significant difference in accuracy rates between favorable and unfavorable sources ($F(1,1122) = 1.75$, $p = 0.19$). Additionally, we found a negative main effect of rating time, indicating that classification accuracy rates decreased from the first to second rating (mean effect = -0.26, $F(1,1158) = 12.19$, $p < .001$), suggesting that phase-2 disrupted participants’ explicit beliefs about source biases (Figure 3b). These effects were not qualified by an interaction between the time of the report and the bias-type on classification accuracy rates ($F(2,1122) = 0.89$, $p = 0.41$).

Notably, source classification accuracy rates depend not only on how well participants discriminate between different sources but also on the degree of bias they manifest in classifying a source as positively or negatively biased (compared to neutral). To dissociate classification-sensitivity from classification-biases we used Signal Detection Theory (SDT), which assumes that source classification is based on a latent “bias strength evidence” variable. For each source this evidence is modeled as a sample from a Gaussian distribution corresponding to the type of this source. Across the three source types, these Gaussians had equal variance but different means, allowing estimation of two discrimination-sensitivities between un/favorable and neutral source-biases (one for favorable and the other for unfavorable). Additionally, the model allowed estimation of two threshold parameters determining classification biases. When evidence falls below the first threshold, the source is classified as unfavorable; between the two thresholds, as neutral; and above the second threshold, as favorable (Figure 3c; see Methods for full specification).

We fitted this model to participant data separately for each of the two classification times (i.e., following Phase-1 of Phase-2), and then regressed, in a linear mixed-effects models, the individual maximum likelihood (ML) sensitivity parameters on classification-time (coded as: first rating = -0.5, second rating = 0.5) and the source’s true bias (coded as: unfavorable = -0.5, favorable = 0.5). This revealed a main effect of rating time on sensitivity ($b = -0.09$, $t(748) = -3.05$, $p = 0.002$), further supporting the idea that biased sources become less distinguishable after completing a task phase wherein the ground truth is unavailable. We found no significant effects

for source's bias nor an interaction (between source's true bias and rating time) with sensitivity (source's bias: b = 0.002, t(748) = 0.04, p = 0.97; interaction effect: b = -0.03, t(748) = -0.34, p = 0.73; see Figure 3d). In similar analyses for the SDT classification-bias parameters we found no significant effects (all p's>.05, Figure 3e).

We next investigated whether individual variability in classification sensitivity was linked to differences in how individuals debias feedback during Phase-2 value-learning. To quantify the overall debias applied to each source-type, for each participant we calculated the mean maximum likelihood (ML) debias parameter, taken from the "free debias" model and corrected for the appropriate direction based on the source-bias type (e.g., the baseline-corrected favorable-feedback debias parameter was averaged with the sign-flipped baseline-corrected unfavorable feedback debias parameter). Higher values on this metric indicate greater debiasing applied to the sources in the correct direction. We found average classification sensitivity from our SDT model correlated significantly across-participants with the overall debias from our RL model (Spearman's correlation; r(186)= 0.29, p < .001; Figure 3f), indicating that debias corrections in the appropriate direction increased with sensitivity in source classification. In contrast, we found no significant correlation between the average classification-bias parameters from our SDT model and overall debias from our RL model (r(186) = 0.04, p = 0.61; Figure S4). Thus, participants who can better discriminate between biased and neutral feedback were also more effective at correcting for feedback biases, showing reduced under-debiasing tendencies.

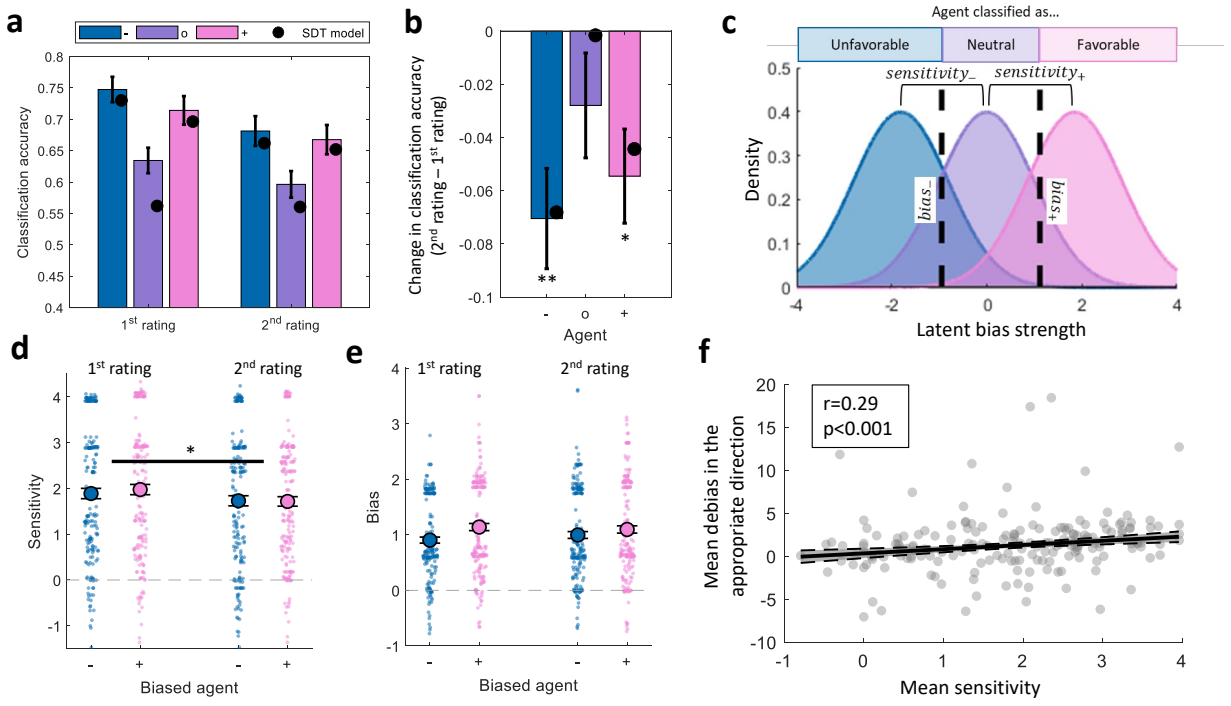

Figure 3: Classification of the bias from each source. (a) Classification accuracy rates as a function of bias type (unfavorable, neutral, or favorable) and classification time (before or after Phase 2). Bars represent participants' accuracy rates, while black circles represent predictions based on maximum-likelihood (ML) parameters from the signal detection theory (SDT) model. **(b)** Change in classification accuracy rates after Phase 2. Participants exhibited a general decrease in classification accuracy rates. **(c)** Signal detection theory (SDT) model. Evidence strength for each source was modeled as a sample for a

Gaussian distribution corresponding to that source-bias type. The 3 source-types Gaussian distributions had equal variance but different means. The neutral source's distribution was centered at 0, while the mean of the favorable and sign-flipped mean for the unfavorable distributions represented the classification sensitivities for each source. Two thresholds (bias- and bias+) determined classification: evidence < bias- was classified as unfavorable, bias- ≤ evidence ≤ bias+ as neutral, and evidence > bias+ as favorable. **(d)** ML sensitivity parameters for favorable and unfavorable sources at each rating time. Sensitivity decreased from the first rating (before Phase 2) to the second (after Phase 2). **(e)** ML bias thresholds for favorable and unfavorable sources at each rating time (note the bias parameter for the unfavorable agency is sign-flipped). **(f)** Correlation between classification sensitivity and feedback- debiasing in the appropriate direction. Classification sensitivity corresponds to the mean ML sensitivity parameter from our SDT model. Debiasing in the appropriate direction is defined as the average debias parameter corrected for the normative debias direction (i.e., flipping the sign of unfavorable debias parameters). Participants with higher mean classification sensitivity for source bias also tended to apply greater debiasing to (Phase-2) feedback from biased sources. Line represents the result of a linear regression on the data, with its s.e.m. (shaded area). Small dots represent individual participants/simulations, while large circles indicate group means. Error bars represent standard errors of the mean (SEM). Unfavorable, neutral, and favorable sources are denoted by the symbols "-" (blue), "o" (violet), and "+" (pink), respectively. (*) $p < .05$, (**) $p < .01$.

Contrastive classifications in the perception of neutral sources

Our signal detection model did not fully capture participants' classification accuracy rates, particularly for unbiased sources (Figure 3a,b), prompting us to examine further how neutral sources are classified. We found evidence for a violation of an important "independence assumption" of the SDT model, according to which classifications are independent across sources. Specifically, we found classification of neutral sources was influenced by a biasing-tendency (i.e., unfavorable or favorable) of their counterpart superblock source (Fig. 4a).

To formally test this, we employed ordinal multinomial logistic regression, regressing the classification of neutral sources (classification categories ordered as unfavorable < neutral < favorable) based on the counterpart super-block source type (unfavorable coded as -0.5, favorable coded as 0.5) and time of rating (after Phase 1 = -0.5, after Phase 2 = 0.5; Figure 4a). This analysis revealed a significant positive interaction between rating time and the counterparts' type ($b = 0.73$, $t(1499) = 3.51$, $p < 0.001$), indicating that the influence of the partner source on neutral-source classifications differed between task phases. On this basis, we also examined the effects of partner source separately for each rating time. During the second rating, partner sources had a significant positive effect ($b = 0.92$, $F(1,1499) = 38.19$, $p < 0.001$), revealing a "contrastive classification effect", whereby participants tended to classify the neutral source bias type as opposite to the partner's bias type (e.g., classified as unfavorable when paired with a favorable source, and vice versa). This effect was absent during the first rating ($b = 0.19$, $F(1,1499) = 1.62$, $p = 0.20$), suggesting the contrastive effect emerged during Phase 2, when participants had access to biased feedback alone (Figure 4b). Participants' confidence ratings for neutral source classification errors revealed a consistent pattern whereby in Phase 2 (as compared with Phase 1) confidence increased for contrastive (compared to non-contrastive) classifications (see Figure S5). Notably, simulations of our SDT model could not reproduce this contrastive classification effect (see Figure S6).

We reasoned that, since counterpart source type influences how neutral sources are perceived and explicitly classified, it should also affect how feedback from neutral sources is debiased. To investigate this, we extended our "free debias" model by introducing separate debiasing

parameters for neutral sources, conditional on whether they were paired with a favorable or unfavorable source. We describe these two neutral-source parameters as the neutral-un/favorable debias parameters respectively (for an un/favorable counterpart). We then estimated individual parameters by fitting this model to the data based on a hypothesis that participants who exhibit stronger explicit contrastive classification (i.e., a greater tendency to classify a neutral source in opposition to its counterpart) would also demonstrate higher levels of "contrastive debiasing." In other words, we expected, they would show greater differences in how they debias feedback from neutral versus counterpart sources during learning.

To quantify the contrastive classification effect, we computed the mean difference between the probability of misclassifying a neutral source as having the opposite bias of its paired source versus misclassifying it as having the same bias. Contrastive debiasing was quantified as the average of the 1) difference between the favorable and the neutral-favorable debias parameters and, 2) the difference between the neutral-unfavorable and unfavorable debias parameters. Supporting our hypothesis, we found a significant positive correlation between contrastive classification and contrastive debiasing (Spearman's correlation, $r(186) = 0.21$, $p < 0.01$) (Figure 4c). In our Discussion we introduce a potential mechanism for how contrastive classification might emerge during Phase 2.

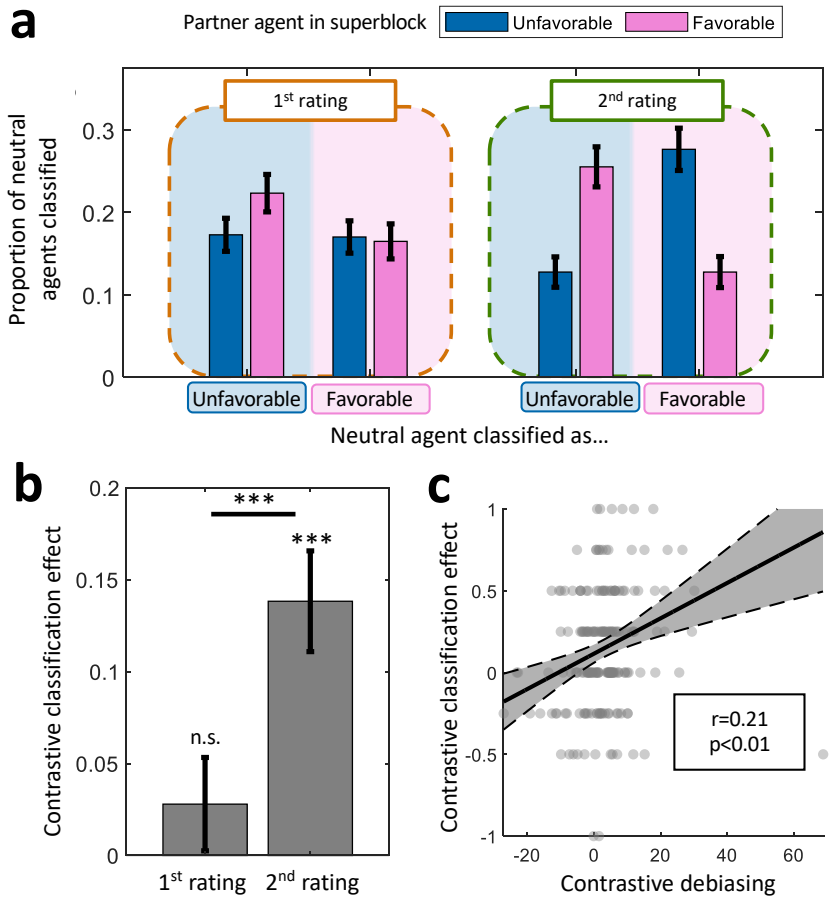

Figure 4: Classification of neutral sources as a function of the other source featured in the superblock. (a) Proportion of neutral sources classified as favorable (pink background) or unfavorable (blue background) based on classification time (before Phase 2, left orange box; after Phase 2, right green box) and the type of the other source featured in the superblock (unfavorable, blue bars; or favorable, pink bars). Neutral sources were more likely to be classified as having the opposite bias to that of the other source in the superblock (i.e., classified as favorable when paired with an unfavorable source, and vice versa), particularly during the second rating. **(b)** Contrastive classification effect as a function of classification time. The contrastive classification effect was calculated as the mean difference between the probability of misclassifying a neutral source as having the opposite bias of its paired source versus misclassifying it as having the same bias. A significant contrastive classification effect was not detected in the first rating, but emerged in the second rating, after participants completed phase 2, where they did not have access to the ground (i.e., unbiased) truth. Error bars represent the standard error of the mean (SEM). **(c)** A scatter plot depicting contrastive classification vs. contrastive debias. Participants who were more likely to classify neutral sources as opposed to their counterparts also exhibited greater divergence in how they debias these source pairs. Line represents the result of a linear regression on the data, with its SEM (shaded area). (***) $p < .001$.

Learning source bias is prioritized over value learning

Our findings show participants leveraged Phase 1 to learn source biases, by comparing source feedback to true outcomes. As we included a bandit choice task in this phase it enables us to examine how individuals trade-off between value learning (for bandits) and bias learning (for sources). Importantly, participants in Phase 1 saw the true outcome before receiving source feedback— a sequence that should support learning bandit values first, followed by learning a source bias (Fig. 1c). However, if bias learning takes precedence, it could deplete cognitive resources needed for effective bandit value learning.

To test this hypothesis, we compared the quality of bandit-choices between phases. Overall, across both phases, choice accuracy rate (i.e., the probability of choosing the painting with the higher value in each offered pair) was significantly above chance (mean accuracy = 0.65, $t(199) = 24.95$, $p < .001$) and improved within each block (average improvement over 24 trials = 0.24, $t(199) = 18.47$, $p < .001$). Notably, accuracy rates were significantly lower in the first compared to the second phase of each superblock (mean difference = 0.065, $t(199) = 9.42$, $p < .001$; Figure 5a). This accuracy difference (between phases) remained stable with task progression i.e., across superblocks (speaking against simple order-effects; Figure 5b; see Figure S8a for detailed results), manifesting especially during the intermediate trials of each block (see Figure S8b).

This result might seem counterintuitive, as participants had access to true choice-outcomes in the first phase alone but had to infer these outcomes based on biased feedback in the second phase (and we know debiasing was somewhat inconsistent across sources; Fig. 2d). However, it turns out the finding is consistent with another possibility, namely that during phase 1, individuals prioritize bias-learning at the expense of value-learning, allowing them to debias feedback during a subsequent phase. To investigate this further, we turned to our (free debias) RL model, which allowed for different learning rates across the two phases of each superblock. Consistent with the idea that value learning is compromised when source biases are learned, the ML learning rate for the first phase was significantly lower than for the second phase (mean difference = 0.26, $t(199) = 12.11$, $p < .001$; Figure 5d). The Phase-2 increase in learning rates predicted, based on model

simulations, an increase in accuracy rates from phase 1 to phase 2 (mean difference = 0.063, $t(199) = 11.37$, $p < .001$; Figure 5c, top panel). By contrast, simulations based on an ablated sub-model, forced to share equal learning rate between both phases, predicted a slight decrease in accuracy rates from the first phase to the second phase (mean difference = -0.006, $t(199) = -2.61$, $p = 0.008$; Figure 5c, bottom panel). Furthermore, individuals who exhibited a greater increase in learning rate from phase 1 to phase 2 also showed a greater improvement in accuracy rates across phases (Spearman's correlation, $r = 0.46$, $p < .001$) (Figure 5e).

If a reduced Phase-1 (bandit-value) learning rate reflects a deployment of greater cognitive resources towards source-bias learning, then participants who exhibited a greater increase in learning from phase 1 to phase 2 should also be better at distinguishing between different source types. To test this, we regressed the average classification sensitivity from our SDT model on both the difference (Phase 2- Phase 1) and sum of the ML learning rates (to control for difference in overall learning of bandit values) for each phase, (see Methods for model specifications). As hypothesized, this revealed a significant effect of learning rate difference on classification sensitivity ($b = 0.35$, $t(184) = 2.12$, $p = 0.035$), indicating participants who showed a larger learning rate increase between phases were also better at classifying the bias levels of different sources (see full coefficients in table S1).

We next reasoned that, during Phase 1, as participants acquire knowledge of source biases, then they should be able to divert available cognitive resources towards bandit value (i.e., they will start using the true outcomes to update bandit values instead of source biases) resulting in an increase in bandit-value learning-rate across trials. Based on this reasoning, we extended our reinforcement learning model to allow learning rates for bandit values to vary dynamically across trials within a block (see Methods for detailed model specifications), computing a learning rate for each phase and trial number (Figure 5e) and regressing these on trial number and phase (Phase 1 = -0.5, Phase 2 = 0.5). Strikingly, we found a significant negative interaction between trial number and phase ($b = -0.02$, $t(9596) = -6.69$, $p < .001$), indicating learning rate dynamics differed between phases. Unpacking this interaction revealed a positive simple effect of trial on learning rate during the first phase ($b = 0.0073$, $F(1,9596) = 18.73$, $p < .001$), consistent with learning rates increasing across Phase 1 trials. Conversely, we observed a negative trial effect on learning rate during the second phase ($b = -0.01$, $F(1,9596) = 31.0$, $p < .001$), reflecting a decline in learning rates during Phase 2. Whereas, our Phase 2 decreasing learning rate finding is consistent with patterns previously described in reinforcement learning literature (21) our Phase 1 findings are consistent with an participants prioritizing source-bias over bandit-value learning, with a gradual resource-shifting towards the latter.

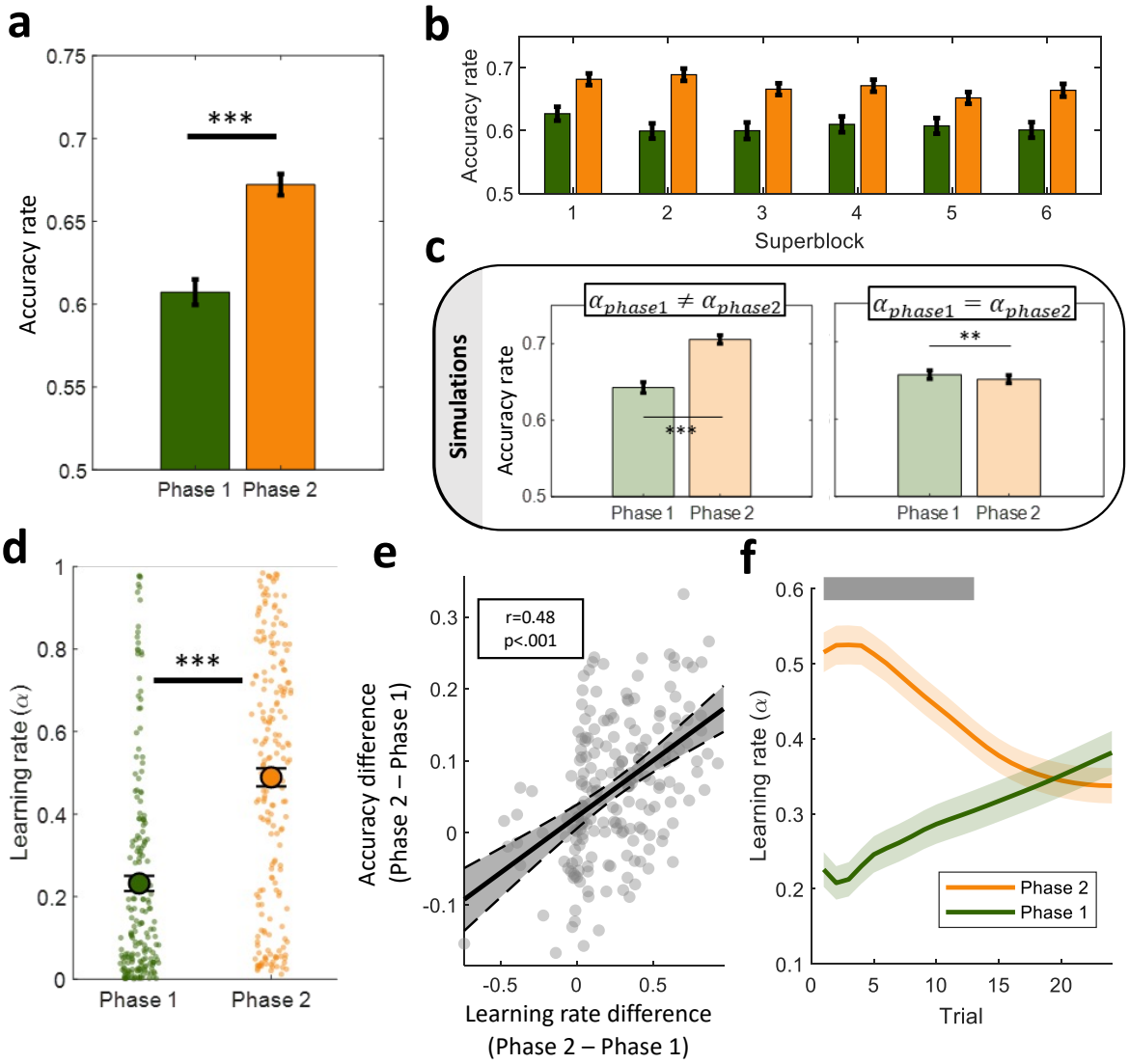

Figure 5: Learning changes from phase 1 (both true outcome and biased feedback visible) to phase 2 (only biased feedback visible). (a) Accuracy rates for each phase of a superblock. Accuracy was higher in Phase 2, where the true outcome was hidden, and participants received potentially biased feedback alone. **(b)** Accuracy rates for each phase plotted separately for each superblock. **(c)** Accuracy rates for simulations based on the ML parameters of participants. The left panel shows simulations based on our free debias model (with separate free learning rate parameters for phase 1 and 2), which predicted an increase in accuracy rates from phase 1 to phase 2. The right panel shows the simulations from a modified version of this model, with a single free learning rate parameter for both phases, predicting a decrease in accuracy rates from phase 1 to phase 2. **(d)** ML learning rate parameters for the first and second phase. Learning rate was higher in Phase 2. **(e)** Scatter plot of changes in ML learning rates vs changes in accuracy rates between the superblock phases. An increase in learning rate going from Phase 1 to Phase 2, was linked to an increase in accuracy rates (i.e., a higher probability of selecting most rewarding bandit within the offered pair). Line represents the result of a linear regression on the data, with its SEM (shaded area). **(f)** Maximum likelihood (ML) learning rate across trials for both phases of the superblock. In Phase 1, participants started with a lower learning rate, which increased over the course of the block. In Phase 2, participants began with a higher learning rate, which decreased as the block progressed. Learning rates were significantly different between phases at the beginning of the block but not at the end. Error bars and

shaded areas represent the standard error of the mean (SEM). Gray boxes highlight significant differences between curves (Bonferroni-corrected for multiple comparisons). (**) $p < .01$,(***) $p < .001$.

Learning the biases of sources does not pay off

Participants prioritized learning source biases in Phase 1 at the expense of value learning, and this led us to next ask whether the ability to effectively debias feedback in Phase 2 compensates for reduced bandit value learning in Phase 1. To address this, we simulated data using our free-debias model (based on ML parameters), which reflects a strategy wherein participants decrease value learning rates in Phase 1 and focus on bias learning, enabling them to debias feedback as a function of the source in Phase 2. We then calculated accuracy rates and reward-rates for these simulated data. Next, we simulated an alternative hypothetical scenario in which participants ignored bias learning in Phase 1, instead learning bandit values as effectively as in Phase 2. However, this came at the cost of applying only a fixed, source-independent, level of debiasing in Phase 2. To model this, we set the Phase 1 learning rate equal to the ML-derived Phase 2 learning rate and fixed all debias parameters to the ML baseline (neutral debias) level.

Our simulations revealed that (compared to hypothetically ignoring source-bias) prioritizing bias learning in Phase 1 significantly reduced the accuracy rate (mean difference = -0.078, $t(199) = -13.14$, $p < 0.001$) and reward acquisition (mean difference = -0.4, $t(199) = -7.38$, $p < 0.001$) during that phase (Figure 6a-b, green bars), without yielding a compensatory improvement in Phase 2 performance (accuracy: mean difference = 0.006, $t(199) = 3.04$, $p = 0.002$; reward acquisition: mean difference = 0.01, $t(199) = 0.42$, $p = 0.67$; Figure 6a-b, orange bars). As a result, overall task performance declined when prioritizing bias learning (accuracy: mean difference = -0.022, $t(199) = -9.26$, $p < 0.001$; reward acquisition: mean difference = -0.09, $t(199) = -3.96$, $p < 0.001$; Figure 6a-b, gray bars).

These findings indicate participants' learning of source biases in Phase 1 led to suboptimal allocation of cognitive resources from a reward-maximization perspective. Additional model-based simulations of hypothetical Phase-2 accuracy rates across a grid of debiasing parameters (Figure 6c; Methods) revealed this inefficiency arose not only because Phase 2 debiasing was partial and inconsistent, but also because, in our task, even perfectly consistent debiasing in Phase 2 provided only a marginal performance benefit (in our discussion, we address alternative conditions where perfect debiasing is more beneficial). While learning to debias feedback may seem intuitively advantageous, the cost of foregoing direct value learning in Phase 1 outweighed potential gains from improved feedback processing in Phase 2.

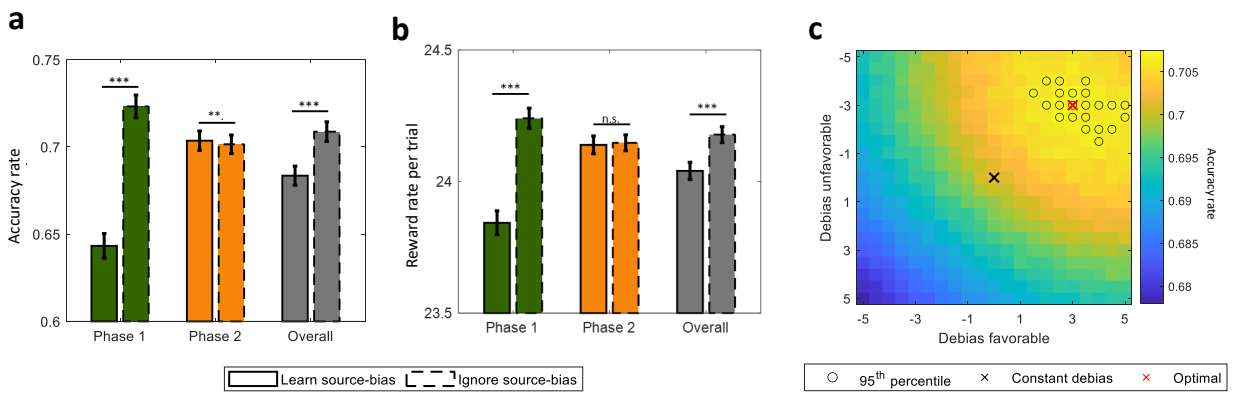

Figure 6: Effects of prioritizing source bias learning at the expense of value learning. (a) Simulated accuracy rates for the strategy used by participants and a hypothetical strategy of focusing on value learning during Phase 1 while ignoring source-bias learning. Solid bars represent accuracy rates in Phase 1 (green), Phase 2 (orange), and the overall task (gray) based on simulations of the "free debias" model using maximum likelihood (ML) parameters. Dashed bars correspond to a hypothetical version of the model, where participants were assumed to ignore bias learning in Phase 1 and focus their learning instead on value-learning (the Phase-1 bandit-value learning rate was fixed to the ML learning rate for Phase 2, and all Phase-2 debias parameters fixed to the ML baseline *neutral-debias* parameter). Hypothetically, ignoring bias learning in Phase 1 would have increased the accuracy rate in Phase 1 while incurring a smaller accuracy rate cost in Phase 2 with an overall accuracy improvement. **(b)** Same as (a) but showing the mean reward rate per trial (true outcomes) acquired instead of accuracy. **(c)** Simulated accuracy rates as a function of the debias applied to unfavorable (y-axis) and favorable (x-axis) agents (compared to the neutral agent baseline). Our simulations were based on our free debias model with participants' ML parameters of participants, with the exception of the debias for biased sources, which were manually adjusted across the presented grid of interest (see Methods). Circled cells represent the 95th percentile of highest accuracies, which falls around the optimal values of -3 (unfavorable) and +3 (favorable) (marked with a red "x") and is marginally better than a constant debias strategy (black "x"). Error bars represent the standard error of the mean (SEM). (**) $p < .01$,(***) $p < .001$.

DISCUSSION

Misinformation research consistently shows individuals are not passive recipients of misinformation but, instead, actively filter it by adjusting their learning processes to reduce a dependence on unreliable sources (20–25). However, this strategy is optimal only when information deviates from the latent truth in unpredictable ways—that is, when misinformation acts as noise (26–28). In contrast, real-world situations often involve information that deviates systematically and predictably from the truth. For instance, news sources are often predictably biased to the left or to the right of the political spectrum. In such cases, individuals can use knowledge about these systematic deviations to reverse-infer a more accurate estimate of the latent truth (33). Indeed, findings in Reinforcement Learning (RL) show that individuals leverage knowledge about the generative structure of the environment to enable causal inference (46) and assign credit appropriately (47–52).

Here, we extend this observation to a biased-misinformation setting to ask how individuals acquire knowledge of source bias and subsequently use this knowledge to correct for biased information (when the ground truth is unavailable). This perspective has been largely overlooked in the misinformation literature, which has primarily focused on source-credibility (23–25), or used

source bias as an indicator of perceived trust (14,32). We found that while participants on average adjust information in a “bias-mitigating” direction, this debiasing was only partial, allowing source bias to influence beliefs. Such imperfect corrections may stem from cognitive limitations (34), such as the cognitive effort required to correct biases, or from inherent cognitive biases like the truth bias (35) or anchoring (36), which lead individuals to accept biased information at face value. It is, however, consistent with other research showing participants struggle to correct (more-complex) biased information (33) and that people continue to rely (somewhat) on misinformation even after it has been successfully debunked (i.e., “continued influence effect”) (53).

These ‘under-corrections’ provide a mechanistic account for the emergence of biased beliefs. Since biased information is not fully corrected, it skews the beliefs of individuals in the direction of the source bias. This means that individuals who start with unbiased beliefs go on to develop biases when repeatedly exposed to systematically biased information (16,54). For example, when a politically biased newspaper criticizes a vaccine's utility, under-debiasing this information will result in unwarranted skepticism. We extend upon previous findings by demonstrating that biased beliefs propagate even when information-source biases are explicitly forewarned and, using a simple additive debiasing strategy, can be easily estimated and corrected. The mechanism we propose is also consistent with evidence that people often share the biases of those they interact with closely, such as family and friends (55–57). We acknowledge a possibility that individuals may have assumed information sources in our task were neutral (35,58), and Phase 1 may not have lasted long enough to fully override these prior beliefs, thereby constraining the extent of debiasing. Future studies could investigate whether extending Phase 1 enhances debiasing by allowing more time for prior beliefs to dissipate.

We show that the very process by which individuals learn about source-bias is itself prone to errors, further impacting an ability to correct misinformation. When ground truth was no longer available we found there was a decline in participants’ classification sensitivity in differentiating between biased sources (i.e., Phase 2). One potential explanation for this is memory decay (59–61) such that, without an anchoring ground truth, participants gradually forget the magnitude or direction of a source’s bias. However, our findings cannot be fully explained by memory decay alone as we observed systematic errors in source classification. After ground-truth feedback was withheld, individuals tended to misclassify information-sources as having a bias opposite to the other information source present in the same superblock. For instance, participants misclassified a neutral source as favorable when paired with an unfavorable source, and vice versa. This contrastive effect also extended to participants’ information-debiasing, such that participants who perceived unbiased and biased sources as having opposing biases subsequently adjusted information from these sources in a more divergent manner. A real-world implication of this is that exposure to left- or right-leaning media may dispose individuals to perceive neutral commentators as actually biased in an opposite direction.

We propose that when ground truth is withheld, an under-debiasing of biased information sources triggers a cascading effect, leading to the contrastive classification of neutral sources. This might arise from competition between potential causes underlying reward-prediction errors (62). To illustrate, when interacting with both neutral and favorable information sources, insufficient debiasing of the latter results in overestimating the value of choice options. As a consequence, neutral feedback will increasingly elicit negative prediction errors that could be attributed either to

choice options (bandits) being less valuable than expected or, alternatively, misattributed to the (objectively) neutral information source being unfavorable. When the latter attribution is used, a contrastive classification emerges. Indeed, we found support for the plausibility of this account in our model simulations (see Figure S7). The observed effect may also be further amplified by a dichotomizing heuristic (63), whereby individuals perceive sources as opposites even in the absence of evidence, or by relative encoding (64,65), where source biases are encoded in relation to the average bias of the surrounding context. Future studies should explore the cognitive mechanisms contributing to contrastive classification and explore avenues that lead to mitigation of its effects. For example, interleaved (rather than blocked) presentation of "ground truth" trials might benefit accurate classifications. Another intriguing hypothesis is that balanced contexts—where favorable and unfavorable sources are equally represented—may not only mitigate the formation of biased beliefs but also foster a more accurate perception of neutral sources.

We hypothesized that when access to ground truth is limited, individuals should capitalize on the few available opportunities to learn about the biases of their information sources. Accordingly, we predicted that participants would allocate substantial cognitive resources to learning these biases when ground truth was accessible (i.e., Phase 1). Our findings strongly support this hypothesis. Counterintuitively, participants learned bandit values at a slower rate when ground truth was available (Phase 1) than when it was not (Phase 2), resulting in lower choice accuracy in Phase 1. This effect persisted despite participants having time to process the true information before receiving source feedback, a sequence that allowed for them to first learn bandit values and subsequently infer source biases (Fig. 2c-d). Furthermore, the Phase 1 decline in bandit-value learning rate predicted participants' sensitivity in explicit source classification. This implies that in a "dual learning problem" (33) learning about source biases is hierarchically prioritized, consuming cognitive resources at the expense of learning about the objects to which the information refers. A real-world example might be when a news story is debunked by fact-checkers, leading readers to revise their perception of the source's reliability rather than update their beliefs about the fact-checked information (66). Furthermore, in the tradeoff between learning about values and source-bias, we found that bandit-value learning rates increased with repeated exposure to ground truth, suggesting that, once participants had learned source biases, they progressively shifted learning-resources to encode bandit values.

A striking finding was that in our task, knowledge of source biases—and the corresponding debiasing feedback—conferred only small benefits to reward rates when ground truth was absent. In this sense, prioritizing source-bias learning over bandit-value learning represented a suboptimal allocation of cognitive resources. We found that hypothetically, participants could have boosted their reward acquisition by disregarding bias learning entirely in Phase 1 and taking feedback at face value in Phase 2. However, despite this potential advantage, we found no evidence that participants reduced their attempts to learn about source biases over time, as the difference in choice accuracy rates between Phases 1 and 2 remained stable throughout the task (Fig. 5b). This suggests a complex trade-off between the costs and benefits of learning about information biases. In some cases, from a purely instrumental perspective, biases in received information might better be ignored, as attempts to counteract them may impose greater costs than the benefits such debiasing affords. It also raises a fundamental question for future research concerning what mechanisms individuals use to determine whether learning about source bias is

worthwhile in the first place. Additionally, it aligns with broader discussions in psychology and neuroscience about resource-rationality (34), particularly how individuals balance the use of a more accurate, but cognitively costly, strategy (e.g., “System 1” or model-based reinforcement learning) with less-accurate but simpler, cost-effective strategies (e.g., “System 2” or model-free learning) (67,68).

Nevertheless, in our task a participants’ tendency to prioritize source-bias learning implies a strong ecological pressure to correct biased feedback. A potential limitation in relation to ecological validity is that source biases were entirely uncorrelated with the bandits. As a result, failing to account for these biases distorted beliefs about all bandits equally, with minimal impact on estimating their relative differences. Since decision quality in our task depended on relative (rather than absolute) bandit values, debiasing provided limited benefits (64,65). In contrast, in real-world settings, biased sources are often context dependent. For example, if a favorable colleague consistently praises one’s research, while an unfavorable colleague critiques one’s teaching, a person who does not correct these biases may come to believe they are a better researcher than teacher, even if the opposite is true. In such cases, accurately learning source biases should significantly enhance decision accuracy (see Figure S9 for a task variant where different agents were associated with different bandits). Furthermore, in many real-world scenarios, success depends on estimating absolute values rather than making relative comparisons (69) —such as when pricing a house for sale. Ultimately, participants may have prioritized source-bias learning in our task because they are accustomed to real-world contexts where it is beneficial, even when it is not the most rational use of cognitive resources in this specific setting (34,70). Prior research suggests that information holds intrinsic value (71,72), which might motivate participants to learn about source biases to infer more accurate (unbiased) feedback or to classify agents based on their biases, even if these classifications were not incentive-compatible and came at the cost of reduced reward acquisition. Despite these considerations, we acknowledge a partial under correction of bias observed in Phase 2 might reflect some participants recognizing that in our task such corrections were relatively unhelpful. Future research should examine whether learning biases are mitigated in contexts where choice objects are correlated with information sources and/or where accurate absolute (rather than relative) beliefs are more strongly incentivized (e.g., by requiring explicit value estimates).

We propose that once biased beliefs of the kind identified here take hold, they may reinforce and recruit additional cognitive biases, such as confirmation bias (31,48,49) or the tendency to trust (27,50) and prioritize (51) information from like-minded individuals. For instance, a motivated cognition account would predict that, when given the choice, participants would preferentially seek out favorable sources—a hypothesis that could be tested in a task variant where participants select their own feedback sources. Furthermore, our model simulations suggest that (partial) misattribution of reward-prediction errors to source bias, combined with confirmatory learning, can trigger a cascading effect. This being the case, even subtle learning biases can snowball into deeply entrenched and exaggerated beliefs over time—affecting both perceptions of bandit values and source credibility (see Figure S10). For example, this could manifest as an increasingly inflated perception of bandit values alongside a progressively decreasing perception of source favorability. In real-world settings, this dynamic might resemble growing skepticism toward

vaccine efficacy, coupled with a perception that information sources are becoming increasingly biased in favor of vaccination.

In summary, we show that while individuals actively attempt to correct for biases in source information, this is often insufficient and results in systematically distorted beliefs. We consider that our paradigm establishes a lower bound on the hazardous impact of biased information on belief updating. In real-world contexts, where biases are more complex (see for example, (33)) and nuanced (e.g., a newspaper may provide accurate financial reporting but exhibit bias on environmental issues) as well as closely intertwined with both the source's and receiver's identity or ideology, these effects are likely to be more pronounced and more resistant to correction. While previous research has largely attributed the societal consequences of misinformation—such as polarization and the formation of echo chambers—to processes of external or self-selection of information (73–75), our findings reveal a potentially more troubling effect. Even when access to ground truth information is provided, biased misinformation can still significantly distort beliefs and decision-making. A key challenge for future research is to refine an understanding of these mechanisms in complex social environments, a necessity if we want to develop effective interventions that mitigate the detrimental effects of misinformation on individual judgment and collective decision-making.

MATERIALS AND METHODS

Participants

We recruited 224 participants (mean age 38.35 ± 12.10, 106 female) from the Prolific participant pool (www.prolific.co) who went on to perform the task on the Gorilla platform (76). 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 3451/001), and all participants provided prior informed consent.

Experimental protocol

Biased information task

Participants were told they would manage an art gallery owning multiple copies of various paintings. They completed six “superblocks”, each consisting of three blocks of a modified multi-armed bandit task, distributed across two phases (the first block in Phase 1 and the second and third blocks in Phase 2). During each block participants chose between four “new” paintings that didn’t appear in any other block. Each round (or trial) of a block involved selecting between two offered paintings (i.e., bandits). A copy of the chosen painting was then sent to an auction house and sold at a variable price based on its quality. For each painting the selling price of copies varied according to a normal distribution, with mean randomly assigned (independently across paintings) to be either $17, $20, $23, $26, or $29, and with a standard deviation of $1. Painting stimuli were created with DALLE-2, a text-to-image AI generator.

Participants were also informed that they would collaborate with different art agencies while making their selections. During each superblock participants interacted with two “new agencies” which were common to all the three super-block blocks but didn’t appear on any of the other superblocks. Each agency was represented as a silhouette with unique art-related icon, which was introduced at the beginning of each superblock and displayed for up to 30 seconds, although participants could proceed earlier if desired.

During each trial participants interacted with one of the two agencies. On every trial, after a painting was selected, an agency provided an estimated selling price for its copy (from here on, “source feedback”), which could differ from the actual, auction-house, selling price (from here on, “true outcome”). Participants were explicitly instructed at the beginning of the study that agencies might tend to either overestimate, underestimate, or offer relatively accurate estimates of a painting’s selling price. However, they were instructed about neither the magnitude of such biases nor the specific estimation tendencies of each of the agencies they encountered during the task. Unbeknownst to participants, each agency belonged to one of 3 “bias types”: favorable agencies provided estimates that were on average $3 above the true selling price, unfavorable agencies $3 below, and realistic agencies around the actual price. Importantly, we tried to remove any underlying motivation from these agencies by instructing participants that agencies might be biased due to past experiences (e.g., some agencies typically deal with more expensive art which may biases them to overestimate prices). Estimates from all agencies had a standard deviation of $3 (i.e., $\text{source feedback} = \text{true outcome} + \mathrm{N}(\mu, 3^2)$, with $\mu = -3, 0, +3$ for unfavorable, neutral and favorable agencies respectively). In Phase 1, samples from these agency bias distributions ($\sim\mathrm{N}(\mu, 3^2)$,) were carefully controlled to ensure that their average exactly matched the normative bias values of the bandits (-3, 0, or +3 for unfavorable, neutral, and favorable agencies, respectively). Within each superblock, the 2 agencies differed in their bias tendency (e.g., if one was favorable the other was either neutral or unfavorable). The three possible pairings of agencies bias-types were counterbalanced across superblocks, with each pairing being featured in two superblocks.

Phase 1 bandit task

During Phase 1 of each superblock, participants completed one block of 24 trials of a modified multi-armed bandit task. In this phase, participants had access to both the true outcome of their choices and the agency feedback, enabling them to form impressions of the biases of the agencies presented in the superblock. Specifically, on each trial, participants were presented with two paintings selected from this set, with all possible painting pairs counterbalanced across the 24 trials (i.e., each bandit pair being presented in 4 trials). The two agencies featured in the superblock were randomly interleaved across trials, with each agency giving feedback in half the trials. Participants were presented with a bandit pair and given a maximum of 2 seconds to make their choice. If participants did not respond on time, they were shown a timeout screen, and proceeded to the next trial. Once a painting was selected (using the left/right key), the unchosen painting disappeared, and the chosen painting was displayed on the screen for 500 ms. Following this, the true outcome of the selected painting was displayed for 2 seconds. The agency responsible for the trial then appeared alongside its feedback for another 2 seconds, while the true outcome remained visible on the screen (see Figure 1c). This setup ensured that participants

could directly compare the true outcomes with the agencies' feedback, facilitating their learning about agency biases.

To maintain consistency and ensure that participants observed the same overall agency biases, the agencies' feedback was carefully controlled. The mean bias of each agency was fixed at -3, 0, or +3 dollars, corresponding to unfavorable, neutral, or favorable agencies, respectively. This control ensured no variation in the overall observed bias within Phase 1.

Phase 2 bandit task

After a 10-second resting period, participants proceeded to Phase 2 which comprised two blocks of 24 trials each. These blocks were similar in structure to Phase 1 but featured a key difference: participants no longer had access to the true outcome of their choices. Since bonus payment depended solely on true outcomes (not agencies' estimates), participants were incentivized to use their Phase-1 knowledge of the agencies' biases to interpret the feedback and accurately estimate the value of each bandit.

Each block included four new bandits (i.e., paintings). Trials worked as in Phase 1 with a main difference: following the presentation of the chosen bandit, the feedback agent for the trial was displayed for 500 ms, after which the agency feedback was shown for 2 seconds (see Figure 1d). Without access to the true outcome, participants were required to use their understanding of the agency's bias to adjust and interpret the feedback effectively.

Additionally, the constraints on the average bias of each agency were relaxed during this phase, allowing bias values to vary slightly from block to block. However, overall, the feedback remained statistically centered around the normative bias levels of -3, 0, or +3 dollars, corresponding to unfavorable, neutral, or favorable agencies, respectively.

Ratings of agency biases

A subset of participants (n = 188) was asked to evaluate their perceptions of agency biases at two points during each superblock: once after Phase 1 (first rating) and once after Phase 2 (second rating). For each rating, participants were presented with each agency individually and asked the question "Overall, do you think the price estimations from this agency tended to be…", and the following response options: "Lower than the true selling price", "Around the true selling price" or "Higher than the true selling price". After clicking on a response, participants provided a confidence rating on a scale from 0 (labeled as "not confident at all") to 100 (labeled as "fully confident"), indicating how confident they were in their assessment.

General protocol

At the beginning of the experiment, participants were presented with instructions for our task. To make the instruction more gradual, participants were first instructed on a classical version of the multi-armed bandit task (excluding the biased feedback agents) and then completed one block of this task. This version worked as the task in phase 1 but omitting the presentation of the agency and its feedback (i.e., participants were only presented with true outcomes). Next, participants received instructions about the biased information task. Upon completing the instructions participants proceeded to the main task. Each wave of instruction (for the classical and the bias

tasks) was interleaved with multiple-choice questions quizzing understanding of the task. 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.

After completing the main task, participants completed four questionnaires (presented in random order): 1) the Patient Health Questionnaire-9 (PHQ-9) (77), 2) the State-Trait Anxiety Inventory for Adults (STAI) (78), 3) the Revised Life Orientation Test (LOT-R) (79), and the Zanarini Rating Scale for Borderline Personality Disorder (80). These questionnaires were used to generate hypotheses for future research.

The participants took on average 59 minutes to complete the experiment. They received a fixed compensation of 8.21 GBP and variable compensation between 0 and 2 GBP based on their performance on the phase 2 of our main task.

Attention checks

The multi-armed bandit tasks in our main task 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 24 participants were excluded, while 200 participants were kept for the analyses.

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. In phase 1, the Q-value of the chosen bandit was updated after reward-feedback based on the true outcome according to a Rescorla-Wagner learning rule:

$$Q \leftarrow Q + \alpha_{phase1} * (TO - Q) \quad \textbf{(3)}$$

where TO is the true outcome of the choice (i.e., the selling price of the selected painting), while α_{phase1} ($\in$[0,1]) is the free parameter representing the learning rate for phase 1. During phase 2, the Q-value of the chosen bandit was updated based on the source feedback, which was debiased conditional on the type of source (unfavorable, neutral or favorable):

$$Q \leftarrow Q + \alpha_{phase2} * (F - debias(source) - Q) \quad (4)$$

where *debias* is a free parameter representing the value used to correct the source feedback (*F*), while α_{phase2} ($\in$[0,1]) is the free parameter representing the learning rate for phase 2. The Q-values for all bandits were initialized with a value of 23, corresponding to the average expected value of the bandits overall.

We formulated three model variants depending on the magnitude of debias of the source feedback. All models included a baseline debias parameter for the neutral source, which represents the correction applied to feedback from neutral sources. However, the treatment of debias parameters for unfavorable and favorable sources varied across the three model variants.

1) In the **constant model**, the debias parameter was not sensitive to the type of source providing feedback, such that $\mathrm{debias(unfavorable)} = \mathrm{debias(neutral)} = \mathrm{debias(favorable)}$. In this case, feedback from all sources was corrected uniformly, without accounting for the specific biases of each source.
2) In the **consistent debias model**, the debias parameters for favorable and unfavorable sources were fixed to their normative (ideal) values relative to the baseline debias for neutral sources: $\mathrm{debias(unfavorable)} = \mathrm{debias(neutral)} - 3$ and $\mathrm{debias(favorable)} = \mathrm{debias(neutral)} + 3$. This model assumes that participants corrected feedback perfectly based on the bias of each source (relative to the baseline debias of the neutral source).
3) The **free debias** model included dedicated free debias parameters, one for each source ($\mathrm{debias(unfavorable)}, \mathrm{debias(neutral)}, \mathrm{debias(favorable)}$), allowing for the possibility the magnitude of corrections could be freely modulated by the type of source bias.

To test the possibility that biased feedback was filtered out through a decrease in learning rate (see section “No evidence for bias effects on learning rate”), we extended our free debias model by including dedicated free learning rate parameters for each source type ($\alpha_{phase2}\mathrm{(unfavorable)}, \alpha_{phase2}\mathrm{(neutral)}, \alpha_{phase2}\mathrm{(favorable)}$).

To test the possibility the debias of neutral agent was influenced by the bias type of the other source present in the superblock, we formulated a **contrastive debiasing model**. This model included separate debias parameters for the neutral agent depending on whether it was paired with a favorable ($debias(neutral|favorable\ partner)$) or an unfavorable ($debias(neutral|unfavorable\ partner)$). In contrast, a single debias parameter was used for both the favorable and unfavorable sources. This formulation allowed the model to capture potential contrastive effects in how participants corrected feedback from neutral sources, depending on the bias of the co-occurring source.

Finally, to test how learning rates evolved during the blocks of each phase, we formulated a “**dynamic learning rate model**”. In this model, the Q-values of bandits were updated as in the free debias model, but with learning rates that dynamically changed across trials within a block. In this model, the learning rate in each phase was modulated by a sigmoid function, allowing it to either increase or decay as a function of the trial number. The trial-specific learning rates for each phase were defined as follows:

$$\alpha_{phase1}(trial) = \frac{1}{1 + e^{-(\eta_{phase1} + \lambda_{phase1} * trial)}} \quad (5)$$

$$\alpha_{phase2}(trial) = \frac{1}{1 + e^{-(\eta_{phase2} + \lambda_{phase2} * trial)}} \quad (6)$$

Here, $\alpha_{phase1}(trial)$ and $\alpha_{phase2}(trial)$ represent the trial-specific learning rates in Phase 1 and Phase 2, respectively. The parameters η_{phase1} and η_{phase2} are used to determine the initial learning rates at the beginning of the block for each phase. The trial number (*trial*) ranges from 0 to 23 within each block, while λ_{phase1} and λ_{phase2} are decay parameters that govern how the learning rate evolves over trials. The sigmoid function allowed the learning rate to follow a smooth, nonlinear trajectory, capturing potential increases or decreases in participants' learning rates as they progressed through a block. Importantly, the separate decay parameters (λ_{phase1} and λ_{phase2}) enabled us to model differences in the dynamics of learning rates between the two phases.

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

$$P(chosen) \leftarrow (1 - f_P) * P(chosen) + PERS \quad (7)$$

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 the two non-shown bandits) were forgotten as follows:

$$P(non - chosen) \leftarrow (1 - f_P) * P(non - chosen) \quad (8)$$

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

$$P(bandit) = \frac{1}{1 + e^{\beta * [Q(other\ bandit) - Q(bandit)] + [P(other\ bandit) - P(bandit)]}} \quad (9)$$

Where β is an inverse temperature parameter applied to the Q-value difference.

Parameter optimization, model selection and synthetic model simulations of RL models

For each participant, we performed maximum a posteriori (MAP) estimation to calculate free parameter values. MAP estimation involves finding the parameter set that maximizes the posterior

probability, which combines the likelihood of the observed data with prior information about the parameters. Specifically, for the inverse temperature parameter β we used a Gamma(1.2, 5.0) prior, and for all other parameters, we applied a Beta(1.1, 1.1) prior transformed to lie within predefined bounds. The bounds for the parameters were as follows: debias ∈ [-20,20], PERS ∈ [-5,5], α ∈ [0,1], f_P ∈ [0,1], λ∈ [-10,10], η∈ [-10,10], β∈ [0,100]. The free parameter values were estimated by maximizing the summed log posterior probability of the observed choices across all games for each participant. Trials where participants exhibited a response time below 150 ms were excluded from the calculations. To minimize the chances of finding local minima, we ran the fitting procedure 10 times for each participant, using random initializations for the parameters (debias∼U[-20,20], PERS∼U[-5,5], α∼U[0,1], f_P∼U[0,1], λ∼U[-10,10]), η∼U[-10,10], β∼U[0,100]).

We performed model-comparisons for nested debias models using generalized-likelihood ratio tests where the null distribution for rejecting a nested model (in favor of a nesting model) was based on a bootstrapping method (BGLRT)(47,81).

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 (1000 total simulations) with a new sequence of trials generated as in the actual data. In Figure 6c and Figure S9b, we used 60 simulations per participant. 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 our free debias model (i.e., our main model of interest), we generated 5 simulations per participants using the ML parameters from our fitting procedure. We fitted each simulated dataset with its generative model and calculated the Spearman's correlation between the generative and fitted parameters.

Signal detection theory model

To model participants' classifications of sources as unfavorable, neutral, or favorable, we applied a Signal Detection Theory (SDT) framework (82). The model assumes that these classifications are based on a latent "bias strength" variable, whose distribution (across same-type sources) depends solely on agent-type (i.e., whether the source is unfavorable, neutral, or favorable). These distributions were Gaussian with equal variance ($\sigma^2 = 1$) but distinct means ($\mu_{unfavorable}$, $\mu_{neutral}$ and $\mu_{favorable}$). The mean of the neutral source distribution was fixed at 0 ($\mu_{neutral} = 0$, without loss of generality, as this is a shifting parameter) , while the means for the unfavorable and favorable sources ($\mu_{unfavorable}$ and $\mu_{favorable}$) were free-parameters estimated from the data. Participants' classifications of evidence depended on two decision thresholds, τ_1 and τ_2 ($\tau_1 < \tau_2$), which defined the boundaries between the three response categories. Specifically, bias strength was classified as unfavorable, neutral or favorable when it was lower than τ_1, between τ_1 and τ_2 or above τ_2, respectively.

The likelihood of each response category was determined by the cumulative probabilities of the Gaussian distributions. For a given type of source, the classification probabilities were computed as follows:

$$P(unfavorable|\mu_{agency},\tau) = \Phi(\tau_1 - \mu_{agency}) \quad (10)$$

$$P(neutral|\mu_{agency},\tau) = \Phi(\tau_2 - \mu_{agency}) - \Phi(\tau_1 - \mu_{agency}) \quad (11)$$

$$P(favorable|\mu_{agency},\tau) = 1 - \Phi(\tau_2 - \mu_{agency}) \quad (12)$$

Here, Φ denotes the cumulative distribution function (CDF) of the standard normal distribution, and μ_{agency} is the mean corresponding to the true type of the classified source ($\mu_{unfavorable}$, $\mu_{neutral}$ and $\mu_{favorable}$).

Finally, we used the maximum likelihood (ML) parameter estimates for each participant to derive sensitivity and bias measures. Sensitivity (d') was calculated for both unfavorable ($d'_{unfavorable} = -\mu_{unfavorable}$) and favorable ($d'_{favorable} = \mu_{favorable}$) sources, reflecting participants' ability to distinguish these sources from neutral ones. Bias was calculated for classifications as unfavorable versus neutral ($bias_{unfavorable} = -\tau_1$) and favorable versus neutral ($bias_{favorable} = \tau_2$). These bias thresholds reflect participants' tendencies to classify evidence into one category over another.

Parameter optimization, model selection and synthetic model simulations of SDT model

As with our RL models, we performed maximum a posteriori (MAP) estimation to calculate the free parameter values for each participant, fitting the two classification phases independently. To ensure that the thresholds fulfilled the condition ($\tau_1 < \tau_2$), we fitted a parameter for τ_1 and a separate parameter for the difference between the thresholds $\Delta\tau$ (i.e. $\tau_2 - \tau_1$), which we forced to be positive. Parameters were bounded as follows: $\mu_{favorable}, \mu_{unfavorable} \in [-100,100]$,$\tau_1 \in [-15,15]$,$\Delta\tau \in [0,30]$. We used a normal prior for all parameters ($\mu_{favorable}, \mu_{unfavorable}, \tau_1, \Delta\tau \sim N(0,10^2)$). The free parameter values were estimated by maximizing the summed log posterior probability of the classifications made by each participant (4 classifications per source). To minimize the chances of finding local minima, we ran the fitting procedure 100 times for each participant, using random initializations for the parameters ($\mu_{favorable}, \mu_{unfavorable} \sim U[-100,100]$,$\tau_1 \sim U[-15,15]$,$\Delta\tau \sim U[0,30]$).

To assess the mechanistic predictions of each model, we calculate the analytic classification probabilities of each source based on the ML parameters of participants.

Mixed effects models

For all our mixed effects models we initially attempted a fit a flexible random effects structure (with all fixed effects being also random effects) but simplified the random effects structure in case of no convergence.

Model-agnostic analysis of source-bias effects on choice-repetition

We used a mixed-effects binomial logistic regression model to assess whether source feedback was perfectly debiased (data from Figure 1a-b). The regressed variable *REPEAT* indicated whether the next trial repeated the choice from the current trial (repeated choice = 1, non-repeated choice = 0), restricted to situations where the next trial featured the chosen bandit (allowing for a choice repetition). Repetition was regressed on the type of source providing feedback in the current trial, encoded in one-hot regressors $\text{AGENT}_{\text{fav.}}$ (current feedback from favorable agent = 1, otherwise = 0) and $\text{AGENT}_{\text{unfav.}}$ (current feedback from unfavorable agent). We also included the regressor *BETTER,* coding whether the bandit chosen on that next trial was the better -higher true mean selling price- or the worse-lower true mean selling price- bandit within the offered pair, coded as 0.5 and -0.5 respectively (and coded as 0 if both bandits had the same mean selling price). Participants served as random effects. The model in Wilkinson's notation was:

$$REPEAT \sim (AGENT_{fav.} + AGENT_{unfav.}) * BETTER + \left((AGENT_{fav.} + AGENT_{unfav.}) * BETTER \middle| participant\right) \quad (13)$$

We also used a mixed-effects binomial logistic regression model to assess whether source feedback was debiased at all (data from Figure 1c). Here, we regressed choice-repetition (*REPEAT*) on $\text{AGENT}_{\text{fav.}}$, $\text{AGENT}_{\text{unfav}}$ and *BETTER* (as in the previous model). Additionally, we included an additional regressor (*FEEDBACK*) encoding the magnitude of the feedback from the agent (centered around 23$). The model in Wilkinson's notation was:

$$REPEAT \sim \left(FEEDBACK + AGENT_{fav.} + AGENT_{unfav.}\right) * BETTER + \left(\left(FEEDBACK + AGENT_{fav.} + AGENT_{unfav.}\right) * BETTER \middle| participant\right) \quad (14)$$

Effects of source-bias on debias parameters

We used a mixed-effects linear regression model to assess whether, and how, debias was modulated by source-bias, with participants serving as random effects (data from Figure 2d). We regressed the maximal likelihood debias parameters from the "free debias" model. The regressors $\text{AGENT}_{\text{fav.}}$ and $\text{AGENT}_{\text{unfav}}$ indicated, respectively, whether the debias parameter was attributed to the unfavorable or favorable agent. The model's Wilkinson's notation was:

$$debias \sim AGENT_{fav.} + AGENT_{unfav.} + (1|participant) \quad (15)$$

Model-agnostic analysis of source-bias effects on classification accuracy

We used a binomial logistic mixed-effects model to test whether the classification accuracy rates of the source bias varied depending on the type of source and the rating time (data from Figure 3a-b). The regressed variable *ACCURACY* indicated whether the source had been correctly classified (correct = 1, incorrect = 0), and was regressed on the following regressors: $AGENT_{fav.}$ (one-hot encoding if the source was favorable), $AGENT_{unfav.}$ (one-hot encoding if the source was unfavorable), and RATING_TIME indicating whether the rating was made before Phase 2 (1st rating = -0.5) or after Phase 2 (2nd rating = 0.5). Participants served as random effects. The model in Wilkinson's notation was:

$$\begin{aligned} ACCURACY \sim\ & RATING\_TIME * (AGENT_{fav.} + AGENT_{unfav.}) \\ & + \left(RATING\_TIME * (AGENT_{fav.} + AGENT_{unfav.}) \middle| participant\right) \end{aligned} \quad (16)$$

Effects of source-bias on classification sensitivity and bias

We used a mixed-effects linear regression model to test whether the sensitivity in classification changed as a function of the type of source and the rating time, with participants serving as random effects (data from Figure 3d-e). We regressed the maximal likelihood sensitivity parameters (d') from the SDT model, on the following regressors: *FAVORABILITY* indicating whether the parameter was attributed to the unfavorable (coded as -.5) or favorable agent (coded as .5), and RATING_TIME indicating whether the rating was made before Phase 2 (1st rating = -0.5) or after Phase 2 (2nd rating = 0.5).. The model's Wilkinson's notation was:

$$\begin{aligned} d' \sim\ & RATING\_TIME * FAVORABILITY \\ & + (RATING\_TIME * FAVORABILITY | participant) \end{aligned} \quad (17)$$

To test for the same effect on classification bias, we repeated the regression but using the ML bias parameters as the regressed variable. The model's Wilkinson's notation was:

$$\begin{aligned} bias \sim\ & RATING\_TIME * FAVORABILITY \\ & + (RATING\_TIME * FAVORABILITY | participant) \end{aligned} \quad (18)$$

Effects of learning rate change between phases on classification sensitivity

We used a linear regression model to test whether individual differences in classification sensitivity depended on the change in learning rate between phases of a superblocks. We regressed the ML sensitivity parameters (d') from the SDT model, on the difference between the

ML learning rate parameters between the phases (*DIFFERENCE*=$\alpha_{phase2} - \alpha_{phase1}$) from our free debias model. We also included the sum of both learning rates as a regressor (*SUM*=$\alpha_{phase2} + \alpha_{phase1}$) to control for individual differences in the overall learning rate (both regressors were centered around their mean across participants). The model's Wilkinson's notation was:

$$d' \sim DIFFERENCE * SUM \quad (19)$$

Effects of phase on learning rate dynamics

We used a mixed effects linear regression model to test whether the changes in learning rate within a block varied depending on the phase of the superblock. We regressed the learning rate for each trial within a block (α), calculated using the ML parameters from our "dynamic learning rate model", on their associated trial (*TRIAL* = [0,23]) and phase (Phase 1 = -0.5; Phase 2 = 0.5), with participants serving as random effects. The model's Wilkinson's notation was:

$$\alpha \sim TRIAL * PHASE + (TRIAL * PHASE | participant) \quad (20)$$

Multinomial logistic regression model

We used an ordinal multinomial logistic regression to test how the classification of the neutral agent (as unfavorable, neutral, or favorable) was influenced by the source type of the partner agent and the time of rating. This approach was implemented using the *fitmnr* function in MATLAB.

The regressed variable *NEUTRAL_CLASSIFICATION* encodes the classification categories (of neutral agents) in an increasing order of favorableness (unfavorable < neutral < favorable). We regressed this variable on *PARTNER,* indicating the true bias type of the counterpart superblock agent (unfavorable=−0.5, favorable = 0.5); and *RATING_TIME*, indicating whether the classification was made after Phase 1 (1st rating, coded as -0.5) or after Phase 2 (2nd rating, coded as 0.5). The model's Wilkinson's notation is:

$$NEUTRAL\_CLASSIFICATION \sim PARTNER * RATING\_TIME \quad (21)$$

ACKNOWLEDGEMENTS

We thank Tali Sharot, Moshe Glickman, Bastien Blain, Mehrdad Salmasi, Jon Rozenbeek, Stefano Palminteri and Peter Dayan 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., and funding from the Max Planck Society to R.J.D. 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.

COMPETING INTERESTS

Authors declare that they have no competing interests.

REFERENCES

1. Roozenbeek J, van der Linden S. The Psychology of Misinformation [Internet]. Cambridge: Cambridge University Press; 2024 [cited 2025 Mar 26]. (Contemporary Social Issues Series). Available from: https://www.cambridge.org/core/books/psychology-of-misinformation/2FF48C2E201E138959A7CF0D01F22D84

2. Horta Ribeiro M, Calais PH, Almeida VAF, Meira W Jr. “Everything I Disagree With is #FakeNews”: Correlating Political Polarization and Spread of Misinformation [Internet]. arXiv e-prints. 2017 [cited 2023 Aug 15]. Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv170605924H

3. Piazza JA. Fake news: the effects of social media disinformation on domestic terrorism. Dyn Asymmetric Confl. 2022 Jan 2;15(1):55–77.

4. Carrieri V, Madio L, Principe F. Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy. Health Econ. 2019;28(11):1377–82.

5. Rocha YM, de Moura GA, Desidério GA, de Oliveira CH, Lourenço FD, de Figueiredo Nicolete LD. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. J Public Health. 2023 Jul 1;31(7):1007–16.

6. Belluz J. Vox. 2017 [cited 2024 Jan 17]. Why Japan’s HPV vaccine rates dropped from 70% to near zero. Available from: https://www.vox.com/science-and-health/2017/12/1/16723912/japan-hpv-vaccine

7. The saga of “Pizzagate”: The fake story that shows how conspiracy theories spread. BBC News [Internet]. 2016 Dec 2 [cited 2024 Jan 17]; Available from: https://www.bbc.com/news/blogs-trending-38156985

8. Enders AM, Uscinski JE, Seelig MI, Klofstad CA, Wuchty S, Funchion JR, et al. The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation. Polit Behav. 2023 Jun 1;45(2):781–804.

9. Pan J, Qi W, Wang Z, Lyu H, Luo J. Bias or Diversity? Unraveling Fine-Grained Thematic Discrepancy in U.S. News Headlines [Internet]. arXiv; 2023 [cited 2024 Dec 11]. Available from: http://arxiv.org/abs/2303.15708

10. Ruan Q, Namee BM, Dong R. Bias Bubbles: Using Semi-Supervised Learning to Measure How Many Biased News Articles Are Around Us.

11. Groseclose T, Milyo J. A Measure of Media Bias*. Q J Econ. 2005 Nov 1;120(4):1191–237.

12. Heseltine M, Clemm Von Hohenberg B, Menchen-Trevino E, Gackowski T, Wojcieszak M. Effects of Over-Time Exposure to Partisan Media and Coverage of Polarization on Perceived Polarization. Polit Commun. 2024 Nov 4;1–22.

13. Weeks BE, Menchen-Trevino E, Calabrese C, Casas A, Wojcieszak M. Partisan media, untrustworthy news sites, and political misperceptions. New Media Soc. 2023 Oct;25(10):2644–62.

14. Schwalbe MC, Joseff K, Woolley S, Cohen GL. When politics trumps truth: Political concordance versus veracity as a determinant of believing, sharing, and recalling the news. J Exp Psychol Gen. 2024;153(10):2524.

15. Garrett RK, Long JA, Jeong MS. From partisan media to misperception: Affective polarization as mediator. J Commun. 2019;69(5):490–512.

16. Glickman M, Sharot T. How human–AI feedback loops alter human perceptual, emotional and social judgements. Nat Hum Behav. 2024 Dec 18;1–15.

17. Glickman M, Sharot T. AI-induced hyper-learning in humans. Curr Opin Psychol. 2024 Dec;60:101900.

18. Bai X, Wang A, Sucholutsky I, Griffiths TL. Explicitly unbiased large language models still form biased associations. Proc Natl Acad Sci. 2025 Feb 25;122(8):e2416228122.

19. Ecker U, Roozenbeek J, van der Linden S, Tay LQ, Cook J, Oreskes N, et al. Misinformation poses a bigger threat to democracy than you might think. Nature. 2024 Jun;630(8015):29–32.

20. Vidal-Perez J, Dolan R, Moran R. Disinformation elicits learning biases. 2024 [cited 2024 Dec 12]; Available from: https://www.researchsquare.com/article/rs-4468218/latest

21. Campbell-Meiklejohn D, Simonsen A, Frith CD, Daw ND. Independent Neural Computation of Value from Other People's Confidence. J Neurosci. 2017 Jan 18;37(3):673–84.

22. De Martino B, Bobadilla-Suarez S, Nouguchi T, Sharot T, Love BC. Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. J Neurosci. 2017 Jun 21;37(25):6066–74.

23. Prike T, Butler LH, Ecker UKH. Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online. Sci Rep. 2024 Mar 22;14(1):6900.

24. Nadarevic L, Reber R, Helmecke AJ, Köse D. Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. Cogn Res Princ Implic. 2020 Nov 11;5:56.

25. Mena P. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. Policy Internet. 2020;12(2):165–83.

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

27. Lee S, Gold JI, Kable JW. The human as delta-rule learner. Decision. 2020;7(1):55–66.

28. Pulcu E, Browning M. The Misestimation of Uncertainty in Affective Disorders. Trends Cogn Sci. 2019 Oct 1;23(10):865–75.

29. Wallace LE, Wegener DT, Petty RE. When Sources Honestly Provide Their Biased Opinion: Bias as a Distinct Source Perception With Independent Effects on Credibility and Persuasion. Pers Soc Psychol Bull. 2020 Mar 1;46(3):439–53.

30. Kahneman D, Sibony O, Sunstein CR. Noise: A flaw in human judgment. Hachette UK; 2021.

31. Baly R, Karadzhov G, Alexandrov D, Glass J, Nakov P. Predicting Factuality of Reporting and Bias of News Media Sources [Internet]. arXiv; 2018 [cited 2024 May 31]. Available from: http://arxiv.org/abs/1810.01765

32. Swire B, Berinsky AJ, Lewandowsky S, Ecker UKH. Processing political misinformation: comprehending the Trump phenomenon. R Soc Open Sci. 2017 Mar;4(3):160802.

33. Schulz L, Schulz E, Bhui R, Dayan P. Mechanisms of Mistrust: A Bayesian Account of Misinformation Learning [Internet]. OSF; 2023 [cited 2024 Jun 14]. Available from: https://osf.io/8egxh

34. Lieder F, Griffiths TL. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behav Brain Sci. 2020;43:e1.

35. Levine TR. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. J Lang Soc Psychol. 2014 Sep 1;33(4):378–92.

36. Furnham A, Boo HC. A literature review of the anchoring effect. J Socio-Econ. 2011 Feb 1;40(1):35–42.

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

38. Lefebvre G, Summerfield C, Bogacz R. A Normative Account of Confirmation Bias During Reinforcement Learning. Neural Comput. 2022 Jan 14;34(2):307–37.

39. Rollwage M, Zmigrod L, de-Wit L, Dolan RJ, Fleming SM. What Underlies Political Polarization? A Manifesto for Computational Political Psychology. Trends Cogn Sci. 2019 Oct 1;23(10):820–2.

40. Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, et al. The psychological drivers of misinformation belief and its resistance to correction. Nat Rev Psychol. 2022 Jan;1(1):13–29.

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

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

43. Globig LK, Holtz N, Sharot T. Changing the Incentive Structure of Social Media Platforms to Halt the Spread of Misinformation. [cited 2022 Oct 13]; Available from: https://psyarxiv.com/26j8w/

44. Lindström B, Bellander M, Schultner DT, Chang A, Tobler PN, Amodio DM. A computational reward learning account of social media engagement. Nat Commun. 2021 Feb 26;12(1):1311.

45. Brady WJ, McLoughlin K, Doan TN, Crockett MJ. How social learning amplifies moral outrage expression in online social networks. Sci Adv. 2021 Aug 13;7(33):eabe5641.

46. Sloman SA, Lagnado D. Causality in Thought. Annu Rev Psychol. 2015 Jan 3;66(Volume 66, 2015):223–47.

47. Moran R, Dayan P, Dolan RJ. Human subjects exploit a cognitive map for credit assignment. Proc Natl Acad Sci. 2021 Jan 26;118(4):e2016884118.

48. Moran R, Keramati M, Dolan RJ. Model based planners reflect on their model-free propensities. PLOS Comput Biol. 2021 Jan 7;17(1):e1008552.

49. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011 Mar 24;69(6):1204–15.

50. Deserno L, Moran R, Michely J, Lee Y, Dayan P, Dolan RJ. Dopamine enhances model-free credit assignment through boosting of retrospective model-based inference. Kahnt T, Büchel C, Cools R, editors. eLife. 2021 Dec 9;10:e67778.

51. Moran R, Dayan P, Dolan RJ. Efficiency and prioritization of inference-based credit assignment. Curr Biol. 2021 Jul 12;31(13):2747-2756.e6.

52. Moran R, Keramati M, Dayan P, Dolan RJ. Retrospective model-based inference guides model-free credit assignment. Nat Commun. 2019 Feb 14;10(1):750.

53. Ecker UKH, Lewandowsky S, Tang DTW. Explicit warnings reduce but do not eliminate the continued influence of misinformation. Mem Cognit. 2010 Dec 1;38(8):1087–100.

54. DellaVigna S, Kaplan E. The Fox News Effect: Media Bias and Voting*. Q J Econ. 2007 Aug 1;122(3):1187–234.

55. Stevenson KT, Peterson MN, Bondell HD. The influence of personal beliefs, friends, and family in building climate change concern among adolescents. Environ Educ Res. 2019 Jun 3;25(6):832–45.

56. Cooperman A. Most U.S. parents pass along their religion and politics to their children [Internet]. Pew Research Center. 2023 [cited 2024 Dec 17]. Available from: https://www.pewresearch.org/short-reads/2023/05/10/most-us-parents-pass-along-their-religion-and-politics-to-their-children/

57. Willoughby EA, Giannelis A, Ludeke S, Klemmensen R, Nørgaard AS, Iacono WG, et al. Parent Contributions to the Development of Political Attitudes in Adoptive and Biological Families. Psychol Sci. 2021 Dec;32(12):2023–34.

58. Brashier NM, Marsh EJ. Judging Truth. Annu Rev Psychol. 2020 Jan 4;71(Volume 71, 2020):499–515.

59. Zhang W, Luck SJ. Sudden Death and Gradual Decay in Visual Working Memory. Psychol Sci. 2009 Apr;20(4):423–8.

60. Brubaker MS, Naveh-Benjamin M. The effects of presentation rate and retention interval on memory for items and associations in younger adults: a simulation of older adults' associative memory deficit. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn. 2014;21(1):1–26.

61. Wang B. Effect of Time Delay on Recognition Memory for Pictures: The Modulatory Role of Emotion. PLoS ONE. 2014 Jun 27;9(6):e100238.

62. Hu L, Kovach M, Li A. Learning News Bias: Misspecifications and Consequences [Internet]. Department of Economics, University of Waterloo; 2023 [cited 2025 Mar 5]. Available from: https://www.uwaterloo.ca/economics/sites/default/files/uploads/documents/learning-news-bias-misspecifications-and-consequences.pdf

63. Kramer HJ, Goldfarb D, Tashjian SM, Hansen Lagattuta K. Dichotomous thinking about social groups: Learning about one group can activate opposite beliefs about another group. Cognit Psychol. 2021 Sep 1;129:101408.

64. Bavard S, Lebreton M, Khamassi M, Coricelli G, Palminteri S. Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. Nat Commun 2018 91. 2018 Oct 29;9(1):1–12.

65. Bavard S, Palminteri S. The functional form of value normalization in human reinforcement learning. Kahnt T, editor. eLife. 2023 Jul 10;12:e83891.

66. Oeldorf-Hirsch A, Schmierbach M, Appelman A, Boyle MP. The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. Mass Commun Soc. 2020 Sep 2;23(5):682–704.

67. Kool W, Cushman FA, Gershman SJ. When Does Model-Based Control Pay Off? PLOS Comput Biol. 2016 Aug 26;12(8):e1005090.

68. Kim D, Park GY, O′Doherty JP, Lee SW. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. Nat Commun. 2019 Dec 16;10(1):5738.

69. Juechems K, Altun T, Hira R, Jarvstad A. Human value learning and representation reflect rational adaptation to task demands. Nat Hum Behav 2022 69. 2022 May 30;6(9):1268–79.

70. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012;35(7):1024–35.

71. Bennett D, Bode S, Brydevall M, Warren H, Murawski C. Intrinsic Valuation of Information in Decision Making under Uncertainty. PLOS Comput Biol. 2016 Jul 14;12(7):e1005020.

72. Bromberg-Martin ES, Monosov IE. Neural circuitry of information seeking. Curr Opin Behav Sci. 2020 Oct;35:62–70.

73. Garrett RK. Echo chambers online?: Politically motivated selective exposure among Internet news users1. J Comput-Mediat Commun. 2009 Jan 1;14(2):265–85.

74. Ross Arguedas A, Robertson C, Fletcher R, Nielsen R. Echo chambers, filter bubbles, and polarisation: a literature review [Internet]. Reuters Institute for the Study of Journalism; 2022 [cited 2024 Apr 29]. Available from: https://ora.ox.ac.uk/objects/uuid:6e357e97-7b16-450a-a827-a92c93729a08

75. Cardenal AS, Aguilar-Paredes C, Galais C, Pérez-Montoro M. Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure. Int J Press. 2019 Oct 1;24(4):465–86.

76. Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. Gorilla in our midst: An online behavioral experiment builder. Behav Res Methods. 2020 Feb 1;52(1):388–407.

77. Kroenke K, Spitzer RL, Williams JBW. Patient Health Questionnaire-9. 1999;

78. Spielberger CD. State-Trait Anxiety Inventory for Adults. 1983;

79. Scheier MF, Carver CS, Bridges MW. Revised Life Orientation Test. 1994;

80. Zanarini MC. Zanarini Rating Scale for Borderline Personality Disorder. 2003;

81. Moran R, Goshen-Gottstein Y. 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. 2015 Jun 1;79:40–67.

82. Macmillan NA, Creelman CD. Detection theory: A user’s guide, 2nd ed. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2005. xix, 492 p. (Detection theory: A user’s guide, 2nd ed).