SUPPLEMEMENTARY INFORMATION

SI Figures

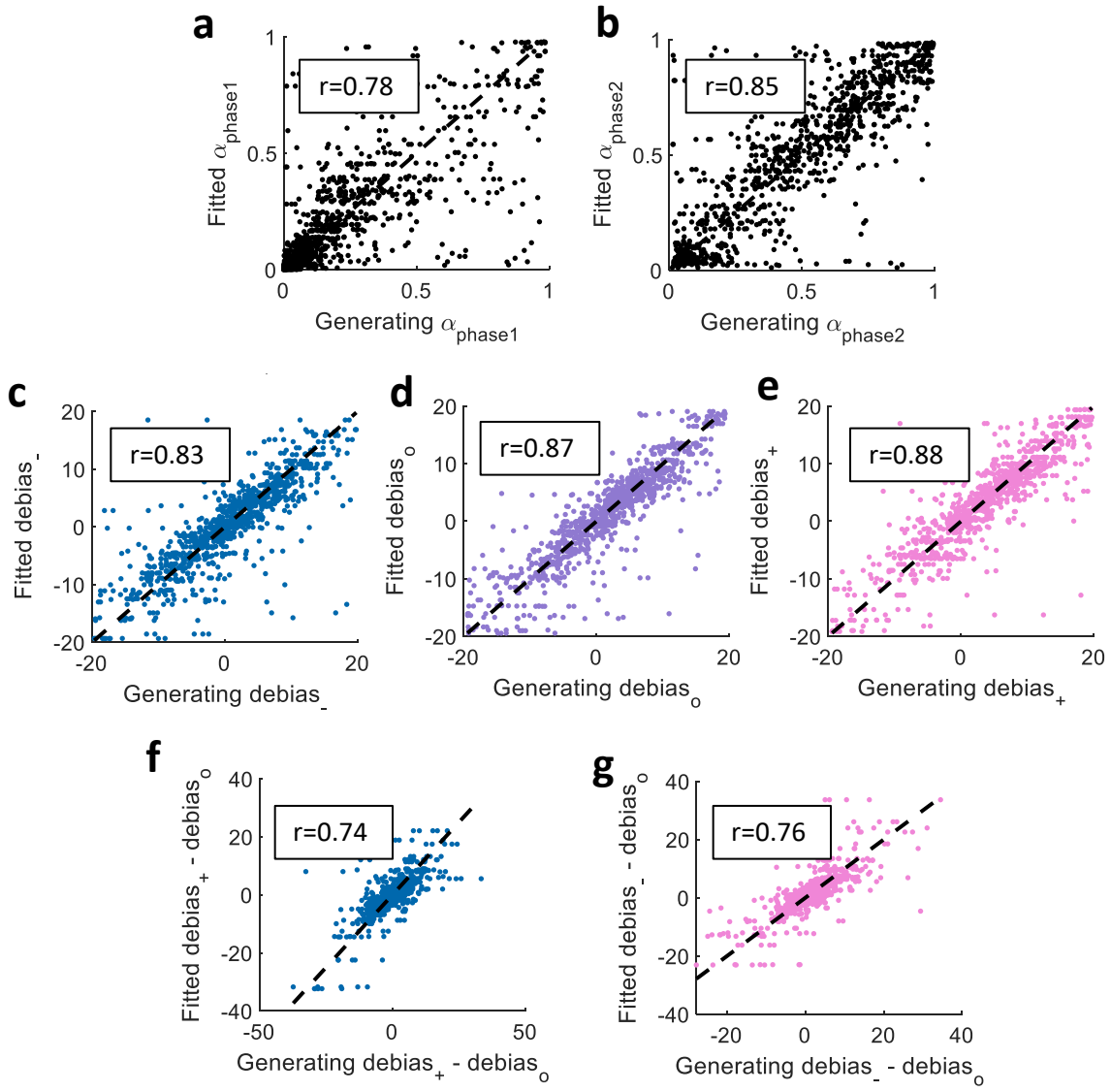

Figure S1: Parameter recovery for parameters of interest from free debias model. Recovery of the learning rate parameter for phase 1 **(a)** and phase 2 **(b),** as well as the debias parameter for the unfavorable **(c)**, neutral **(d)** and favorable **(e)** feedback, and for the debias parameters for unfavorable **(f)** and favorable **(g)** feedback relative to neutral feedback. We generated 5 simulations per participants using the ML parameters from our "free debias" model fitting procedure (generating parameters; x-axes), and we then fitted each simulated dataset with again its "free debias" model (fitted parameters; y-axes). Recoverability was defined as the Spearman correlation between generating and fitted parameters. Circles represent the

datapoints of individual simulations (1000 per parameter; 5 per participant), the denoted metric "r" corresponds to the Spearman correlation between the generative and fitted parameters.

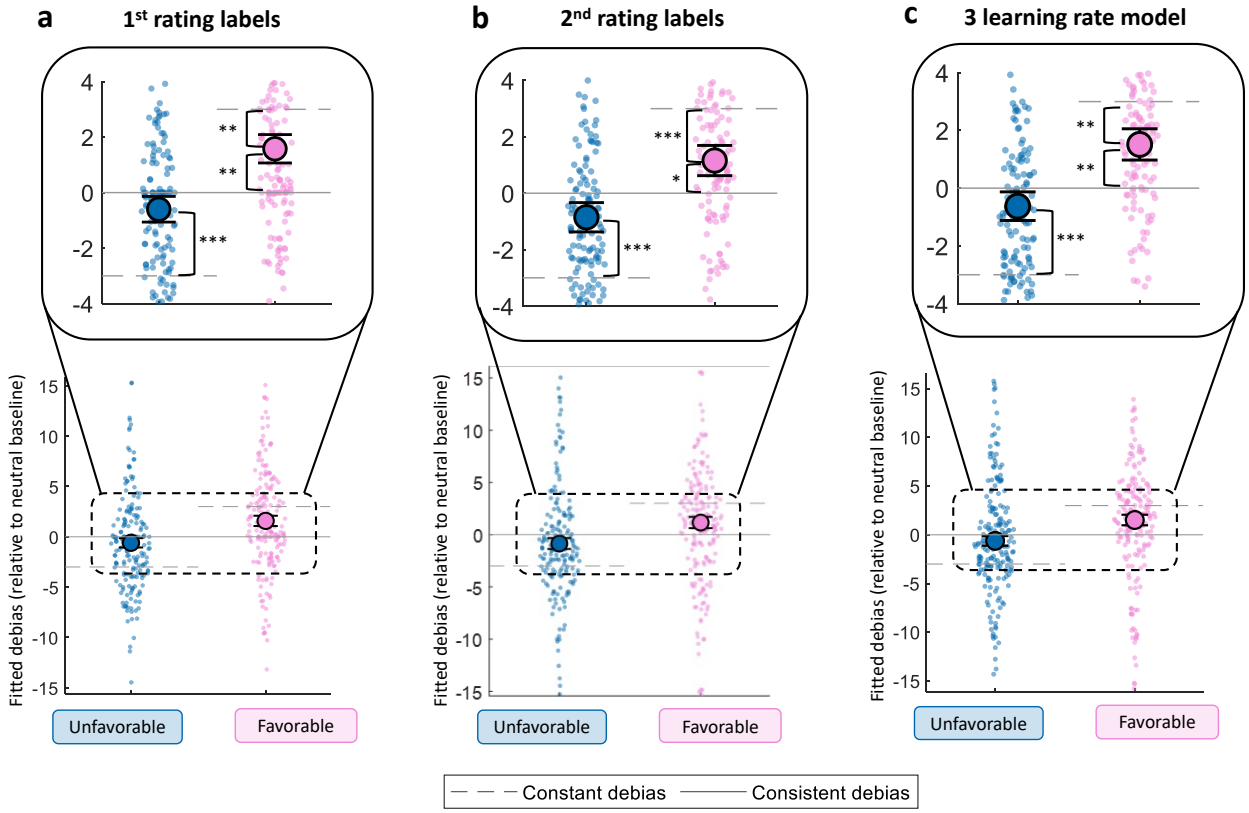

Figure S2. ML estimates of debias parameters for favorable and unfavorable feedback relative to neutral feedback for alternative models. (a) ML parameters for variant of free debias model where biased sources are labeled based on the participants classifications during the first rating (irrespective of their true type). **(b)** Same as previous plot but based on the classifications from the second rating. **(c)** ML parameters from extension of free debias model dedicated learning rates for each source in phase 2. 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. (**) $p < .01$,(***) $p < .001$.

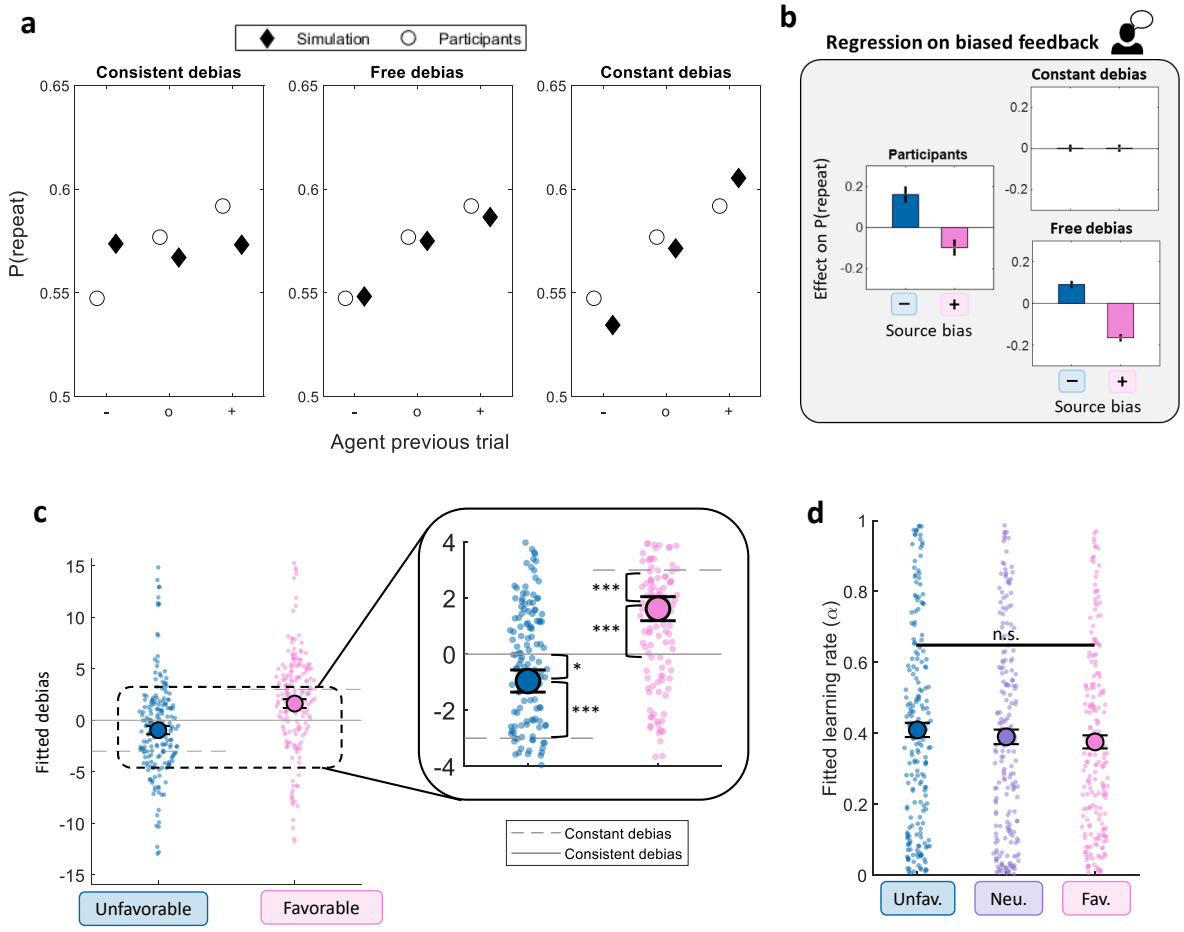

Figure S3. Characterization of source debiasing for "0-debias" variants of the models, where the baseline (i.e., neutral-source) bias was fixed to 0. (a) Probability of repeating a choice (when offered on two consecutive trials) as a function of the source providing feedback in the previous trial for the different model versions. The "constant" model (right) assumes no correction of source bias (i.e., feedback is taken at face value); the "consistent debias" model (middle) fully corrects feedback bias by applying a debias of +3, 0, and -3 to feedback from unfavorable, neutral, and favorable sources, respectively; and the "free debias" model (bottom) treats debiasing of the unfavorable and favorable agencies as free parameters (with the debias of the neutral agency fixed to 0). **(b)** Regression analysis of choice repetition based on biased feedback and the type of source. **(c)** ML estimates of debias parameters for favorable and unfavorable feedback based on our free 0-debias model (with neutral debias set to 0). 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 (with neutral debias fixed to 0). 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. (*) $p < .05$, (***) $p < .001$, (n.s.) $p>.05$.

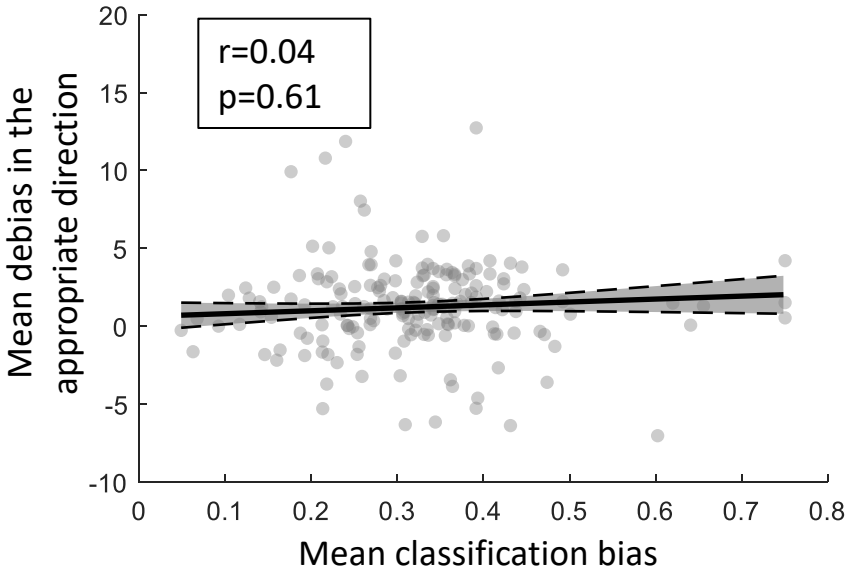

Figure S4. Correlation between classification sensitivity and debiasing in the appropriate direction. Classification bias corresponds to the mean ML bias parameter from our SDT model. Debiasing in the appropriate direction is defined as the average debias parameter corrected for the normative debias direction (i.e., flipping the sign of unfavorable debias parameters). 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).

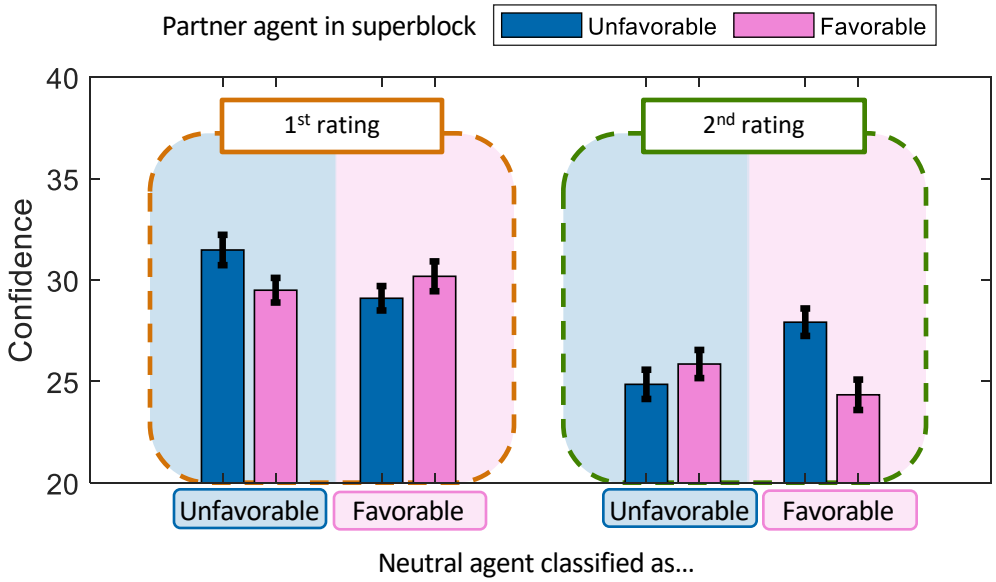

Figure S5. Contrastive classification in reported classification confidence. Confidence in the classification of neutral sources either classified as favorable (pink background) or unfavorable (blue background) based on classification time (before Phase 2, left orange box; after Phase 2, right green box) and the type of the other agent featured in the superblock (unfavorable, blue bars; or favorable, pink bars). We regressed, in a linear mixed effects model, the classification confidence on the misclassification

response (neutral classified as unfavorable = -0.5; classified as favorable = 0.5), the partner source (unfavorable = -0.5, favorable = 0.5) and the rating time (first rating = -0.5, second rating = 0.5). This analysis revealed a significant triple interaction between misclassification response, rating time and the partner source type (b = -7.81, t(472) = -2.84, p = 0.004). During the first rating, we found a marginally significant positive interaction between misclassification response and partner sources (b = 3.71, F(1,472) = 3.21 p = 0.07), suggesting that participants tended to show higher confidence when misclassifying a neutral source as having the same bias-type as the partner. In contrast, we found a marginally significant negative interaction in the second rating (b = -4.09, F(1,472) = 3.56, p = 0.06), such that participants were more confident when misclassifying a neutral source as having the opposite bias-type to the partner source (akin to the contrastive effect shown at the level of classification accuracy). Error bars represent the standard error of the mean (SEM).

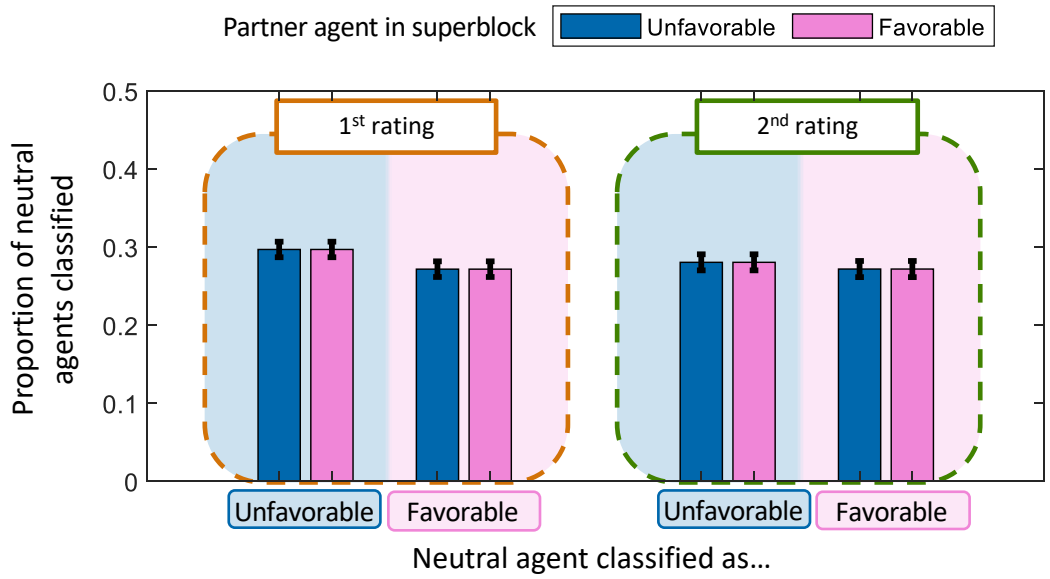

Figure S6. SDT model predictions of classification of neutral agents as a function of the other agent featured in the superblock. Proportion of neutral agents classified as favorable (pink background) or unfavorable (blue background) based on classification time (before Phase 2, left orange box; after Phase 2, right green box) and the type of the other agent featured in the superblock (unfavorable, blue bars; or favorable, pink bars). The same ordinal multinomial logistic regression from the main text did not reveal a significant interaction between rating time and the counterparts' type (p = 1), nor a main effect of the counterpart's type (p = 1). Consequently, our SDT model did not predict the contrastive classification effect that we found in participants. Error bars represent the standard error of the mean (SEM).

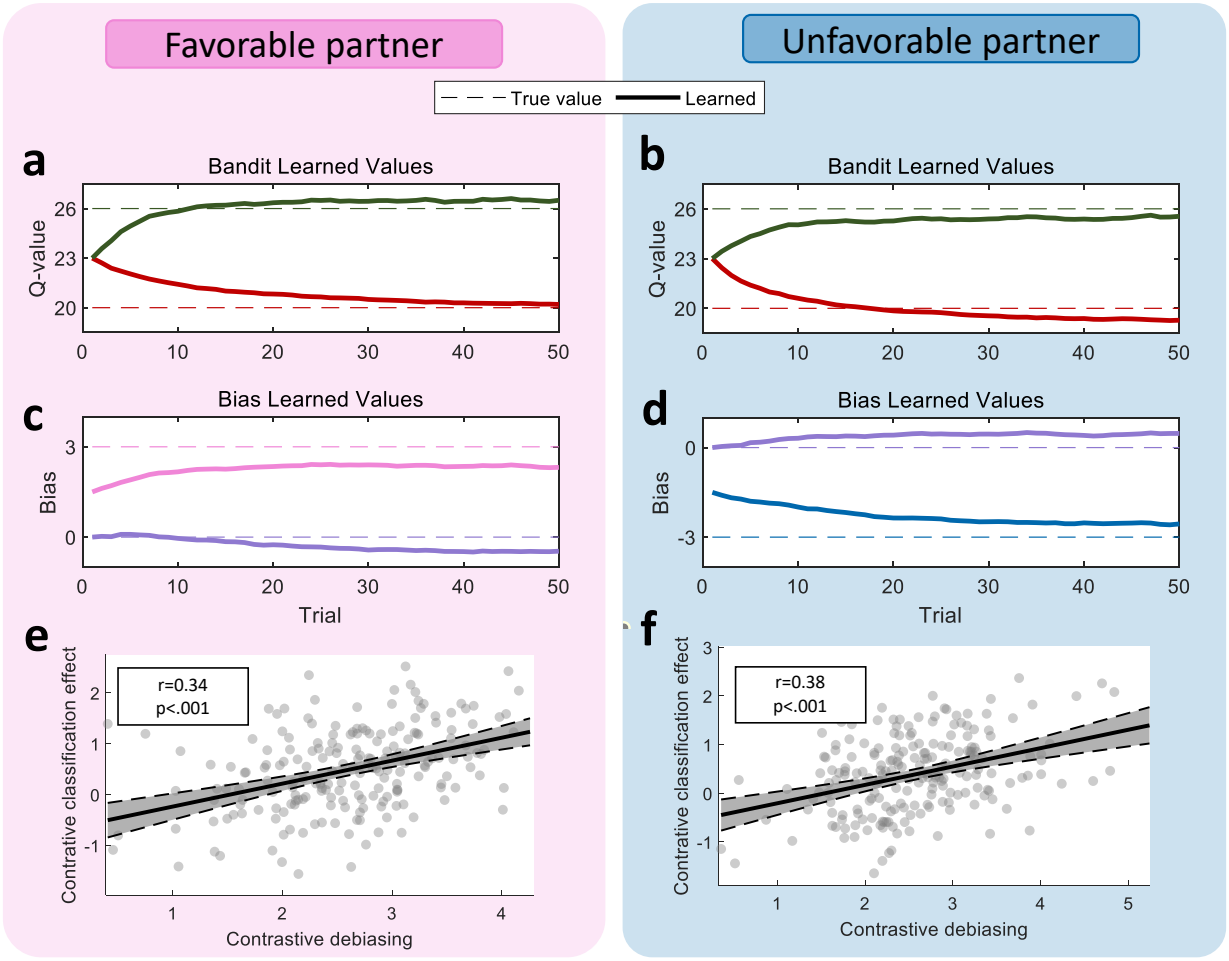

Figure S7: Simulations illustrating a suggested mechanism underlying the emergence of the contrastive classification effect for neutral sources. We simulated Phase 2 of a simplified task (with only two bandits) under the assumption that debias parameters learned in Phase 1 were below optimal levels (1.5 for favorable sources, -1.5 for unfavorable sources). In this simulation, participants relied on Phase 2 reward prediction errors to update their beliefs about both bandit values and source biases, which in turn were used to debias feedback from different sources (see SI Methods for full model details). **(a)** Evolution of Q-values (solid lines) relative to true expected values (dashed lines) in a favorable-neutral context (i.e., when one information source is neutral and the other is favorable). Underdebiasing of the favorable source inflates Q-values above their true expected values. **(b)** Same as (a), but in an unfavorable-neutral context. Underdebiasing of the unfavorable source deflates Q-values below their true expected values. **(c)** Evolution of perceived biases for neutral (purple solid line) and favorable (pink solid line) sources relative to their ideal values (dashed lines) in a favorable-neutral context. Since underdebiasing of the favorable source leads to inflated Q-values, feedback from a neutral source appears lower than expected, progressively contributing to a perception this source is negatively biased. **(d)** Same as (c), but in an unfavorable-neutral context. Here, underdebiasing of the unfavorable source results in deflated Q-values, so neutral feedback appears more positive than expected, contributing to a perception this source is favorable **(e)** Scatter plot depicting the relationship between contrastive classification and contrastive debiasing in the favorable-neutral context. For contrastive classification, we used as a proxy the signed-flipped belief about the (objectively) neutral-source bias at the end of the simulation ($-\boldsymbol{neutral\ bias}\ (\boldsymbol{t} = \boldsymbol{50})$). For, contrastive debiasing we took as a proxy the mean difference between the favorable and neutral biases across each simulation ($\langle \boldsymbol{favorable\ bias}(\boldsymbol{t}) \rangle - \langle \boldsymbol{neutral\ bias}(\boldsymbol{t}) \rangle$). Simulations where neutral sources were perceived as more unfavorable at the end of the phase also exhibited greater difference between bias across source types throughout the block, mirroring the pattern observed in participants (see Figure 4c in main text). **(f)** Same as (e), but for simulations in an unfavorable-neutral context. Here, our

proxies were contrastive classification: ($\boldsymbol{neutral\ bias}\ (\boldsymbol{t} = \mathbf{50})$), contrastive debiasing: ($\langle \boldsymbol{neutral\ bias}(\boldsymbol{t}) \rangle - \langle \boldsymbol{unfavorable\ bias}(\boldsymbol{t}) \rangle$). Dots represent individual simulation results, while the solid line represents a linear regression fit with its SEM (shaded area).

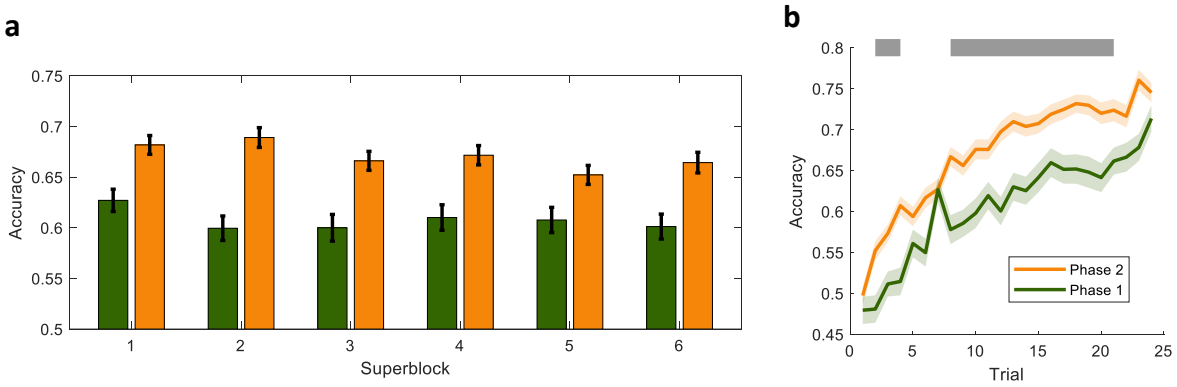

Figure S8: Learning changes from phase 1 (both true outcome and biased feedback available) to phase 2 (only biased feedback available). (a) Accuracy rates for each phase plotted separately for each superblock. We used a binomial logistic mixed-effects model to regress accuracy rates on the superblock phase (Phase 1 = -0.5; Phase 2 = 0.5) and superblock number (1,2,...,6 then centralized). This revealed main effects for phase (b = 0.29, t(2388) = 16.92, p < .001) and for superblock number (b = -0.019, t(2388) = -3.89, p < .001), but not of their interaction (b = -0.007, t(2388) = -0.70, p = 0.48), suggesting a consistently higher phase 2 accuracy across superblocks. **(b)** Accuracy curves for the blocks in phase 1 and phase 2. Differences in accuracy across phases were especially high for trials in the middle of the block. 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).

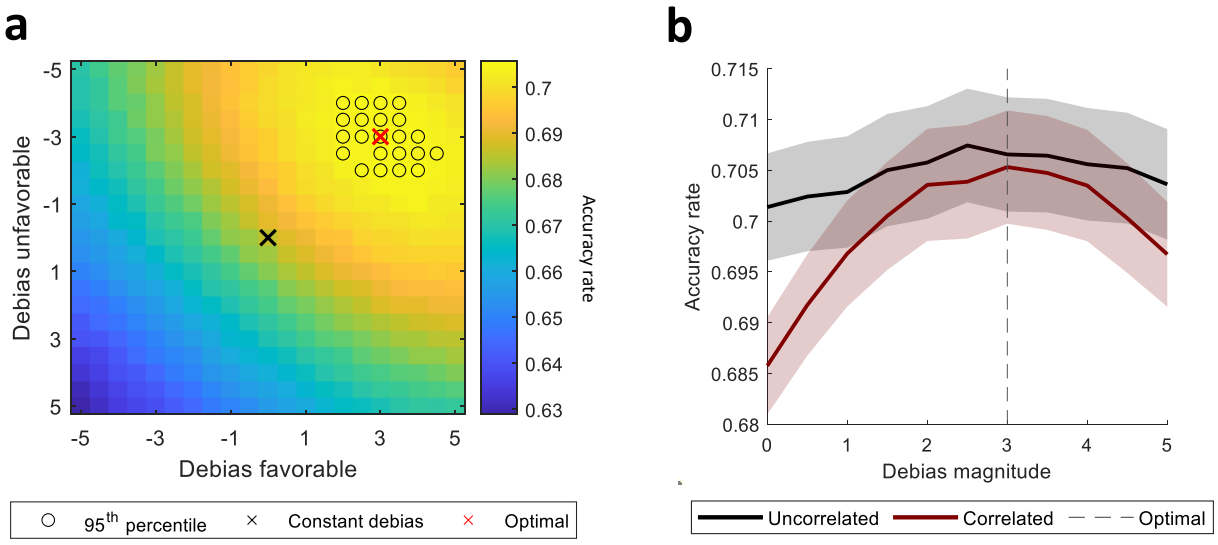

Figure S9: Effects of debiasing accuracy on task where each biased source is associated with different bandits. We simulated an alternative version of our task where agents and bandits were fully correlated. In this alternative version of the task each bandit was associated with a given source, such that feedback about that bandit would only be provided by that source (each agent in a superblock always gave feedback for two out of the four bandits in a given block). **(a)** Accuracy rates as a function of the debias

applied to unfavorable (y-axis) and favorable (x-axis) agents (compared to the neutral agent baseline). Accuracy rates were based simulations of our free debias model in our current task. Simulations were based on the ML parameters of participants (with the exception of the debias for biased sources, which was our variable of interest). We performed 60 simulations per participants and averaged across them. Circled cells represent the 95th percentile of highest accuracies (marked with black circles), which falls around the optimal values of -3 (unfavorable) and +3 (favorable) (marked with red "x"), showing an accuracy improvement when compared with a constant debias strategy (black "x"). **(b)** Comparison of impact of debias sensitivity on accuracy for our current task (bandits and agents uncorrelated; black line), and the alternative task version (bandits and agents correlated; red line). Debias magnitude is defined as the correction added to unfavorable feedback and subtracted from favorable feedback (relative to the neutral baseline). Both curves peaked around the ideal value of 3, but deviations from such value had a greater impact on accuracy when biased sources were correlated with bandits. Shaded areas represent the standard error of the mean (SEM).

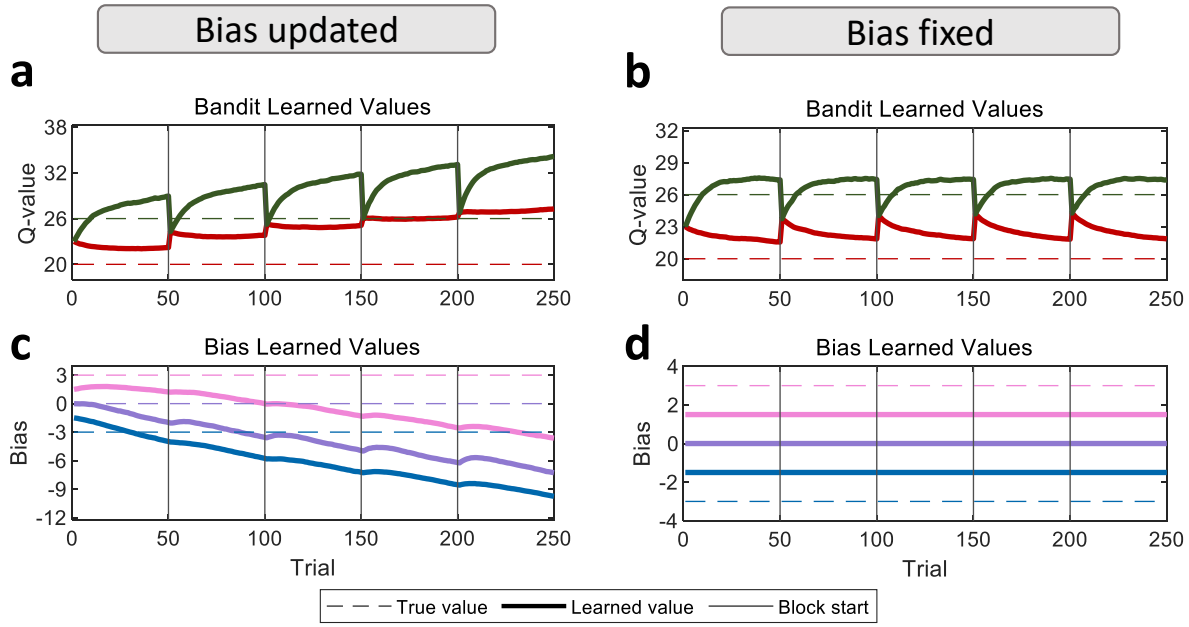

Figure S10: Updating the estimated bias of sources can lead to a snowball inflation of beliefs. We simulated a simplified version of Phase 2 (6 blocks of 50 trials each; new pair of bandits per block) with the 3 source types randomly interleaved throughout each block. In this simulation, participants relied on Phase-2 reward prediction errors to update their beliefs about both bandit-values and about source-bias. Beliefs about source bias were used in turn to debias source-feedback. Additionally, our simulation model assumed: (1) a positivity bias in learning bandit values, where positive prediction errors (PEs) contribute more to learning than negative PEs, and (2) That at the beginning of each block initial bandit values (Q0) are set to the mean value of all bandits encountered previously (see SI Methods a-b for model details). **(a)** Simulations from a model in which participants continuously update their beliefs about source biases. The positivity bias in learning leads to a persistent inflation of perceived bandit values across blocks (top panel). **(c)** The inflation beliefs about bandit values (in panels a) are explained by progressively escalating beliefs about agents being unfavorable. **(b,d)** Same as (a,c), but for a model where bias of sources are not updated. This model does not predict the snowball increase in the value of bandits.

SI Tables

Table 1: mixed effects linear regression model regressing classification sensitivity (ML d' parameters) on the difference and sum of ML learning rate parameters across phases ($\alpha_{phase2} - \alpha_{phase1}$ and $\alpha_{phase2} + \alpha_{phase1}$ respectively), as well as their interaction. The main regressors were centered around their mean across participants. These results are discussed in the section "Learning source bias is prioritized over value learning" in the main text.

Regressor	Coefficient	tStat	DF	p-value
SUM	0.35	2.12	184	0.035
DIFFERENCE	-0.30	-2.68	184	0.008
INTERACTION	0.77	-1.80	184	0.074

SI Methods

a) Source bias updating model

In figures S7 and S10, we formulated a new model to test if the contrastive classification effect could emerge from a combination of underdebiasing and a continuous update of the perceived bias of sources during Phase 2. This model operates similar to our free debias model (excluding the perseveration term), but dynamically updates estimated bias source (i.e., the debias parameter for the source) during Phase 2 as follows:

$$bias(source) \leftarrow bias(source) + \lambda * (F(source) - bias - Q(chosen)) \quad (1)$$

Where the estimated bias for a give source is updated based on the Q-value of the chosen bandit and the feedback about that bandit from the source (*F(source)*), with a learning rate λ. In other words, this model uses reward-prediction errors to update beliefs about both bandits and source-bias.

Bandit Q-values were updated as in our free debias model:

$$Q \leftarrow Q + \alpha * (F(source) - bias(source) - Q) \quad (2)$$

b) Simulations of mechanism explaining emergence of the contrastive classification effect (figure S7)

To show how the contrastive classification effect could emerge from a combination of underdebiasing and continuous source bias updates (Figure S7), we computed 200 simulations of the "source bias updating model" in a simplified version of the Phase 2 of our task. This version included 50 trials featuring two bandits (expected values of 20 and 26) and two biased sources (either neutral and favorable, or neutral and unfavorable, randomly interleaved). Bandits and

agents operated as in our main task. Q-values for the bandits were initialized to 23, and biases were initialized to half their ideal value (1.5 for favorable sources and -1.5 for unfavorable ones). The parameters in the model were fixed to the following values: $\alpha = 0.3, \beta = 0.3, \lambda = 0.1$.

We checked if the emergent contrastive classification effect correlated with the difference in biases across agents in a context (as in Figure 4c in the main text). We used the neutral bias at the end of the block (signed flipped for favorable-neutral context) as a proxy for contrastive classification. We used the mean difference between biases across the simulation as a measure of contrastive debias (favorable-neutral in favorable context, and neutral-unfavorable in unfavorable context). We computed the Spearman's correlation for the 200 simulated datasets.

c) Simulation of snowball effect in the perceived bias of bandits (figure S10)

In order to show that Phase 2 updating of biases (as in the previous section) in combination with positivity bias could lead to a snowball effect (Figure S10), we extended the "source bias updating model" by allowing a positivity bias in the update of bandit values. The new learning rule for bandit values was therefore:

$$\begin{aligned} \delta &= F - debias(agency) - Q \\ Q &\leftarrow Q + \alpha^{+} * \delta \quad for\ \delta > 0 \\ Q &\leftarrow Q + \alpha^{-} * \delta \quad for\ \delta < 0 \end{aligned} \tag{1}$$

where positive and negative prediction errors (δ) contribute differently to learning, as specified by the different learning rate parameters (α^{+} for positive PEs, and α^{-} for negative PEs).

We computed 400 simulations of this model. Each simulation encompassed 5 blocks (of 50 trials each), with each block featuring 2 new bandits (with values 20 and 26). For the first block we initialized Q-values to 23, but for every other block we initialize them to the mean Q-values at the end of all previous blocks. The positive and negative learning rates for these simulations were set to 0.3 and 0.1, respectively. All other parameters and task variables were the same as in the previous section.