Supplementary Information

Niccolò Pescetelli[1, 2] and Nick Yeung[1]

[1]Department of Experimental Psychology, University of Oxford
[2]Max Planck Institute for Human Development, Berlin, Germany

August 9, 2020

1 Supplementary methods

1.1 Experiment 1: Advisors description

After the participant confirmed their initial perceptual decision response with the spacebar, one of four different advisors appeared centrally as a head-shot picture. The advisors pictures were all Caucasian, smiling female characters (Tottenham et al., 2009), randomly assigned per participant to the four accuracy/calibration conditions described below. Advice was provided in the form of spoken sentences (2 s long), that expressed a binary level of confidence (low vs. high) and either agreement or disagreement with the participant's judgment. Low confidence was expressed by the sentences "I think it was on the [LEFT/RIGHT]" and "It was on the [LEFT/RIGHT], I think", with one of the two versions randomly assigned on every trial. Similarly, high confidence was expressed by the sentences "I'm sure it was on the [LEFT/RIGHT]!" and "It was on the [LEFT/RIGHT], I'm sure!". The use of two inverted sentences for each confidence cue ("I'm sure" vs. "I think") was to avoid over-repetition of a single sentence and to balance the differences in emphasis that the English language conveys when using the confidence cues at the beginning or end of the sentence. The selection of LEFT or RIGHT depended on the advisor's choice and accuracy as described below. The spoken advice was pre-recorded from four female native English speakers, again randomised to conditions across participants.

Advisor calibration was defined as the strength of co-variation between confidence judgments and accuracy, and quantified as Type 2 A_{ROC} (A''_{ROC}), a method which that not make assumptions about the generative model of confidence (Fleming & Lau, 2014). Uncalibrated advisors both had an A''_{ROC} of 0.5, meaning that confidence was totally uninformative in predicting the advisor's trial-level accuracy. Due to the experimental design—in which overall accuracy of advisors was fixed at 60% or 80%, and calibrated advisors were always correct when high in confidence—calibrated advisors differed in their metacognitive sensitivity according to this metric.

1.2 Advice Value as Information Gain

Experiment 1 We formalised an advisor's informational value as the mean absolute information gained after each possible social encounter with a specific advisor. Information gain is the difference between the posterior and prior probability of participant's correct response:

$$IG = p(d = w|e) - p(d = w) \tag{1}$$

where posterior probability correct $p(d = w|e)$ represents the probability that the decision d is equal to the correct decision w, conditional on the specific social encounter e. Social encounter e represents one of the four possible events: the advisor (1) confidently disagrees, (2) unconfidently disagrees, (3) unconfidently agrees, (4) confidently agrees (where "confidently" and "unconfidently" refer to the level of confidence expressed by the advisor on that trial). Posterior probability $p(d = w|e)$ was computed using Bayes' theorem and was proportional to participant's prior probability correct $p(d = w)$ and the likelihood of the social event given participant's accuracy $p(e|d = w)$. Given the staircase procedure, we used 70% as prior $p(d = w)$. The probability of agreement (or disagreement) conditional on correct response and the overall probability of agreement (or disagreement) were known by design. The mean absolute information gain so computed was lowest for the Inaccurate Uncalibrated advisor (0.08), intermediate for the Accurate Calibrated and Accurate Uncalibrated advisors (0.29 and 0.26 respectively) and highest for the Calibrated but Inaccurate advisor (0.38). This can be intuitively understood by looking at Table 1, in the main text. Although the Inaccurate Calibrated advisor's accuracy rate is lower than the Accurate Calibrated advisor, outcomes can be better predicted by its judgment. In particular, its judgments correlate strongly positively when sure and strongly negatively when unsure with the correct answer. On the contrary when the Accurate Calibrated advisor is unsure there is a much higher uncertainty about the final outcome. We also computed an expected information gain IG_e for each advisor (Table 1 in the main text) by scaling IG by the overall probability of each event:

$$IG_e = IG * p(e) \tag{2}$$

where $p(e)$ is the overall probability of each social event (i.e., confident disagreement, unconfident disagreement, unconfident agreement, confident agreement). The expected information gain captures the idea that extremely informative but very unlikely events are not very valuable. IG_e values for each advisor were: Accurate Calibrated = .063, Accurate Uncalibrated = .063, Inaccurate Calibrated = .084, Inaccurate Uncalibrated = .021; suggesting that the Inaccurate Calibrated advisor's advice was the most informative.

Experiment 2 Similar to Experiment 1, we used conditional probabilities and the participants' expected accuracy to compute the informational value of each advisor. Advisors' mean absolute information gain IG and expected information gain IG_e were computed as in the previous experiment. Contrary to Experiment 1, however, advisors did not express different levels of confidence. This created only two possible social situations e on each trial (instead of four as in the previous experiment), namely either agreement or disagreement. Information gain was highest for the accurate advisors (.28 and .27 for the high and low agreement advisors respectively) and the lowest for inaccurate advisors (.03 and .06 for the high and low agreement advisors respectively).

Experiment 3 The intuition that the anti-bias advisor was the most informative of the three advisors designed for Experiment 3 was confirmed using a numerical simulation.

We used numerical simulation rather than analytic calculations here because, in Experiment 3, the profile of the advisors cannot be calculated *a priori* but is dependent on the specific distribution of confidence of the participant. The simulations were based on an ideal Bayesian observer performing the task, with a Gaussian distribution of confidence centred on 25 and with a standard deviation of 10. For each initial confidence judgment, the information gained from observing agreement or disagreement was computed for each advisor as the difference between posterior confidence and prior confidence. Contrary to previous Experiments, prior probability correct was here defined on a trial level based on pre-advice confidence. An expected information gain was computed by multiplying the information gain so obtained by the normalisation term in the Bayes formula. This produced a curve of expected information gains after agreeing or disagreeing with each advisor over possible pre-advice confidence levels (Figure S1). The average area under the expected information gain curve was taken as an objective measure of advisor informativeness. Average areas under the curve were 14.74 for the unbiased advisor, 14.63 for the bias-sharing advisor and 15.78 for the anti-bias advisor. This procedure thus quantified and confirmed the intuition that the anti-bias advisor provided the most informative advice.

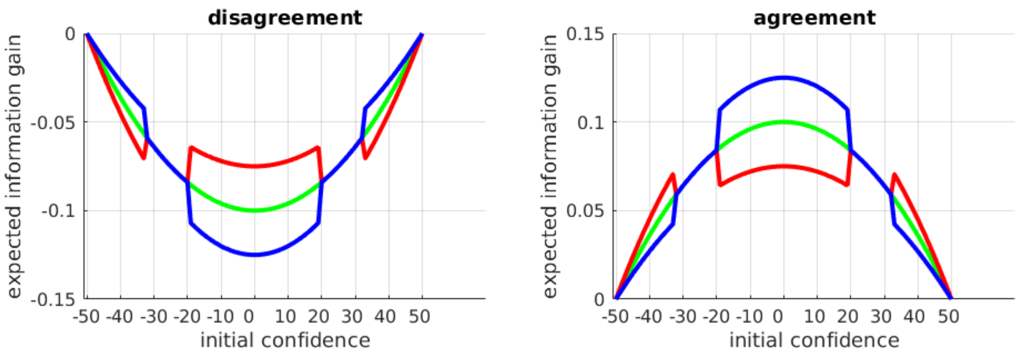

Figure S1: Information gained after agreeing and disagreeing with each advisor type (color code) for each initial subjective prior confidence. Information gain is scaled by the likelihood of agreement and disagreement events. Advice informativeness can be quantified by the area under the curve.

	Experiment 1 Advisors			
	Accurate Calibrated	Accurate Uncalibrated	Inaccurate Calibrated	Inaccurate Uncalibrated
A''_{ROC}	.72	.5	0.84	.5
IG	0.29	0.26	0.38	0.08
$\mathrm{IG_e}$	0.063	0.063	0.084	0.021
	Experiment 2 Advisors			
	High Accuracy High Agreement	High Acc. Low Agr.	Low Acc. High Agr.	Low Acc. Low Agr.
IG	0.28	0.27	0.03	0.06
IG_e	0.09	0.13	0.01	0.03
	Advisors			
	Bias-sharing	Unbiased	Anti-bias	
$AUC(IG_e)$	14.63	14.74	15.78	

Table S1: Experiment 1-3 information gain and expected information gain—IG and IG_e respectively—indicate average informational value of the advice, computed as information gain and expected information gain respectively. Experiment 1 also shows advisors' calibration, measured as type II AROC.

1.3 Measures of interest

Two measures of estimated advice reliability were defined. The first was the explicit trust that participants expressed in the advisors as collected by the brief questionnaires presented to participants every two blocks. In Experiment 1, four questions asked participants to directly rate on a scale from 1 ("Not at all") to 50 ("Extremely") how much they thought each advisor was accurate (Q1), confident (Q2a), trustworthy (Q3) and influential on their own choices (Q4). In Experiments 2 and 3, due to the absence of a confidence judgment from advisors, question 2 was replaced with a question asking about how much participants liked each advisor (Q2b: likeability question). For all Experiments, the first questionnaire was presented immediately after the practice blocks but before any interaction with the advisors took place so to provide a baseline measure. Baseline ratings were subtracted from following ratings to account for confounding factors related to advisors' appearance and inter-individual differences in the use of the scale. A principal component analysis (PCA) was performed for dimensionality reduction on normalised difference scores and the first component was taken as a unitary measure of expressed trust.

In Experiment 1, question loadings for the Feedback condition were 0.52 (Q1), 0.44 (Q2a), 0.50 (Q3), 0.52 (Q4); and for the No-Feedback condition were 0.51 (Q1), 0.39 (Q2), 0.54 (Q3), 0.52 (Q4).

In Experiment 2, question loadings for the Feedback condition were 0.53 (Q1), 0.39 (Q2b), 0.53 (Q3), 0.52 (Q4); and for the No-Feedback condition were 0.51 (Q1), 0.41 (Q2), 0.53 (Q3), 0.51 (Q4).

In Experiment 3, question loadings for the Feedback condition were 0.53 (Q1), 0.42 (Q2b), 0.53 (Q3), 0.50 (Q4); and for the No-Feedback condition were 0.48 (Q1), 0.47

(Q2), 0.53 (Q3), 0.50 (Q4).

The second measure of interest was an implicit index of advisor's influence on participant's opinions, quantifying participants' confidence change from pre- to post-advice:

$$\delta_C = C_{post} - C_{pre}. \tag{3}$$

where C_{pre} is, for Experiment 1 and 2, an integer value between +1 and +5 and C_{post} is an integer value between -5 and +5 (negative C_{post} representing changes of mind). Positive δ_C values mean increases in confidence from pre- to post-advice and negative values represent decreases in confidence. Notice that δ_C values have a negative skew, ranging from -10 (moving from highest confidence in one judgment to highest confidence in the opposite judgment) to +4 (moving from lowest to highest confidence rating for a single judgment). In Experiment 3, given the difference scale used, C_{post} can assume values between -50 and 50, while δ_C can range from -100 to 49. Agreement and disagreement trials typically have opposite effects on confidence change: agreement usually leads to increases in confidence while disagreement to confidence decreases. The absolute magnitude of confidence shifts in both agreement and disagreement trials can be expected to grow larger as the participant makes more use of the advice received. Thus a unitary measure of influence was obtained by subtracting average δ_C in disagreement from average δ_C in agreement:

$$I = \bar{\delta}_C^a - \bar{\delta}_C^d \tag{4}$$

where I assumes greater values as participant's confidence increases in agreement and confidence decreases in disagreement become larger.

2 Supplementary Results

2.1 Confidence change by agreement

Previous research has shown that confidence inversely predicts advice taking and that people discount disagreeing advice. The crucial aim of our investigation is to show that confidence and agreement not only influence how each piece of agreeing/disagreeing advice is weighted, but are also aggregated across interactions to discern something about the advisors themselves (i.e., their overall reliability). This aim was reflected in the definition of our influence measure, which assesses pre- to post-advice changes in confidence separately for trials with agreeing vs. disagreeing advice from each advisor, and calculates influence as the difference between these changes. Thus, via our influence measure, we can show that advisors who more regularly disagree with a participants' choices are less influential, even on those trials where their advice happens to agree with the participants' view. Conversely, we can show that advisors who more regularly agree with the participant are more influential on those trials where their opinion diverges from the participants' initial choice. To further explore these effects, and provide further evidence that the observed influence differences reflect gradually-learned evaluations of advisors' relative reliability, we repeated the analysis and plots reported in the main text, but now assessing confidence change as a function of agreement, separately for each advisor. If people are simply discounting disagreeing advice (even if proportionally to their initial confidence) we would not expect to see differences across advisors, because all agreement

(/disagreement) trials should count the same, independently of who is agreeing. Instead if people are forming a durable representation of others' competence, we would expect to observe differences in confidence change when people agree (disagree) with different advisors. In particular, we would expect a greater influence (i.e., confidence change) of advisors that are believed to be more more competent.

2.2 Experiment 1

We replicate the results observed for the aggregate influence measure (Equation 3) also for agreement and disagreement trials separately. Figure S2 shows the confidence change observed for each advisor, broken down by agreeing and disagreeing trials. As expected, accurate and calibrated advisors produced larger confidence changes in both agreement and disagreement trials.

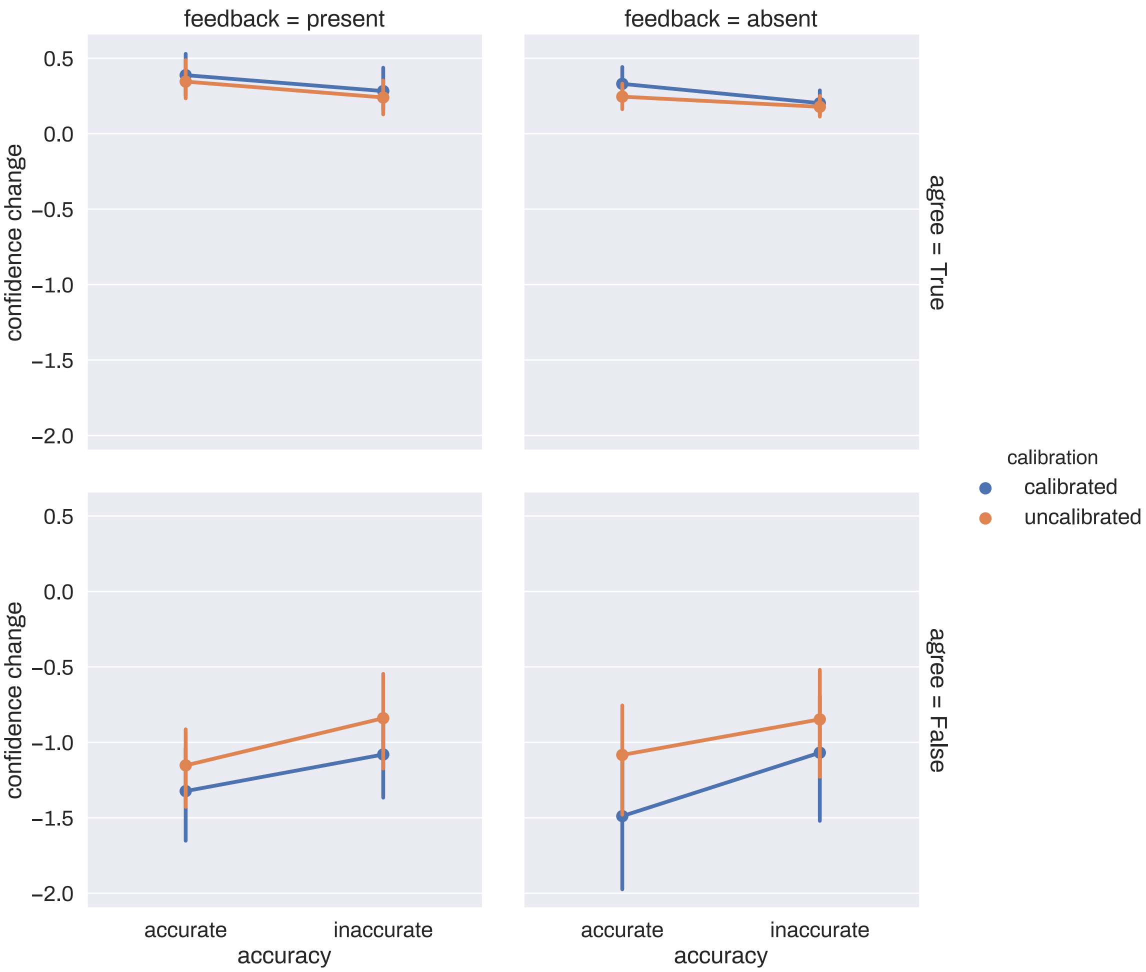

Figure S2: Experiment 1 - Confidence change in agreement and disagreement trials for different advisors. We observe an effect of both advisor's accuracy and calibration (see table S2) in both agreement and disagreement trials. This suggests that people accumulated a representation of others' competence irrespective, instead of simply discounting disagreeing evidence.

Effect	F(1,44)	p	ges
Agreement			
fb	0.75731824	3.888921e-01	1.082965e-02
acc	9.22407505	4.006146e-03**	1.969010e-02
cal	4.00170962	5.164876e-02.	4.576533e-03
conf	69.29001445	1.388161e-10***	1.311394e-01
fb:acc	0.01367809	9.074292e-01	2.978340e-05
fb:cal	0.06240406	8.038984e-01	7.169104e-05
fb:oc	4.65711595	3.642327e-02 *	1.004259e-02
acc:cal	0.41823365	5.211792e-01	4.502113e-04
acc:conf	0.43401602	5.134579e-01	3.094162e-04
cal:conf	3.37609059	7.290780e-02 .	2.128078e-03
fb:acc:cal	0.41563349	5.224712e-01	4.474136e-04
fb:acc:conf	0.17938831	6.739627e-01	1.279117e-04
fb:cal:conf	2.98407362	9.110176e-02 .	1.881440e-03
acc:cal:conf	0.74141098	3.938773e-01	2.551690e-04
fb:acc:cal:conf	1.83713762	1.822040e-01	6.320432e-04
Disagreement			
fb	0.006121536	9.379917e-01	1.011952e-04
acc	15.304270849	3.133341e-04 **	1.725678e-02
cal	18.823806847	8.256714e-05 ***	1.270240e-02
conf	38.697718603	1.598035e-07 ***	8.451080e-02
fb:acc	0.105297640	7.471003e-01	1.208017e-04
fb:cal	0.812641129	3.722487e-01	5.551212e-04
fb:conf	0.377914265	5.418879e-01	9.006905e-04
acc:cal	0.164358350	6.871398e-01	1.563503e-04
acc:conf	2.652761665	1.105102e-01	1.045018e-03
cal:conf	7.753482060	7.874907e-03 *	3.359669e-03
fb:acc:cal	0.814506124	3.717056e-01	7.743419e-04
fb:acc:conf	0.572707598	4.532184e-01	2.257951e-04
fb:cal:conf	0.189169288	6.657348e-01	8.223868e-05
acc:cal:conf	4.049781660	5.032305e-02 .	8.000137e-04
fb:acc:cal:conf	2.531817122	1.187317e-01	5.002976e-04

Table S2: Experiment 1 - confidence change broken down by agreement. We find similar results reported for influence reported in the main text, namely main effects for advisor's accuracy and calibration. Columns from left to right: Effect, F statistic (numerator's degrees of freedom, denominator's degrees of freedom), p-value, generalized η^2_G measure of effect size. Effect abbreviations: fb (feedback), acc (advisor accuracy), cal (advisor calibration), conf (advisor confidence). Significance values: . (< .1), * (< .05), ** (< .01), *** (< .001).

2.3 Experiment 2

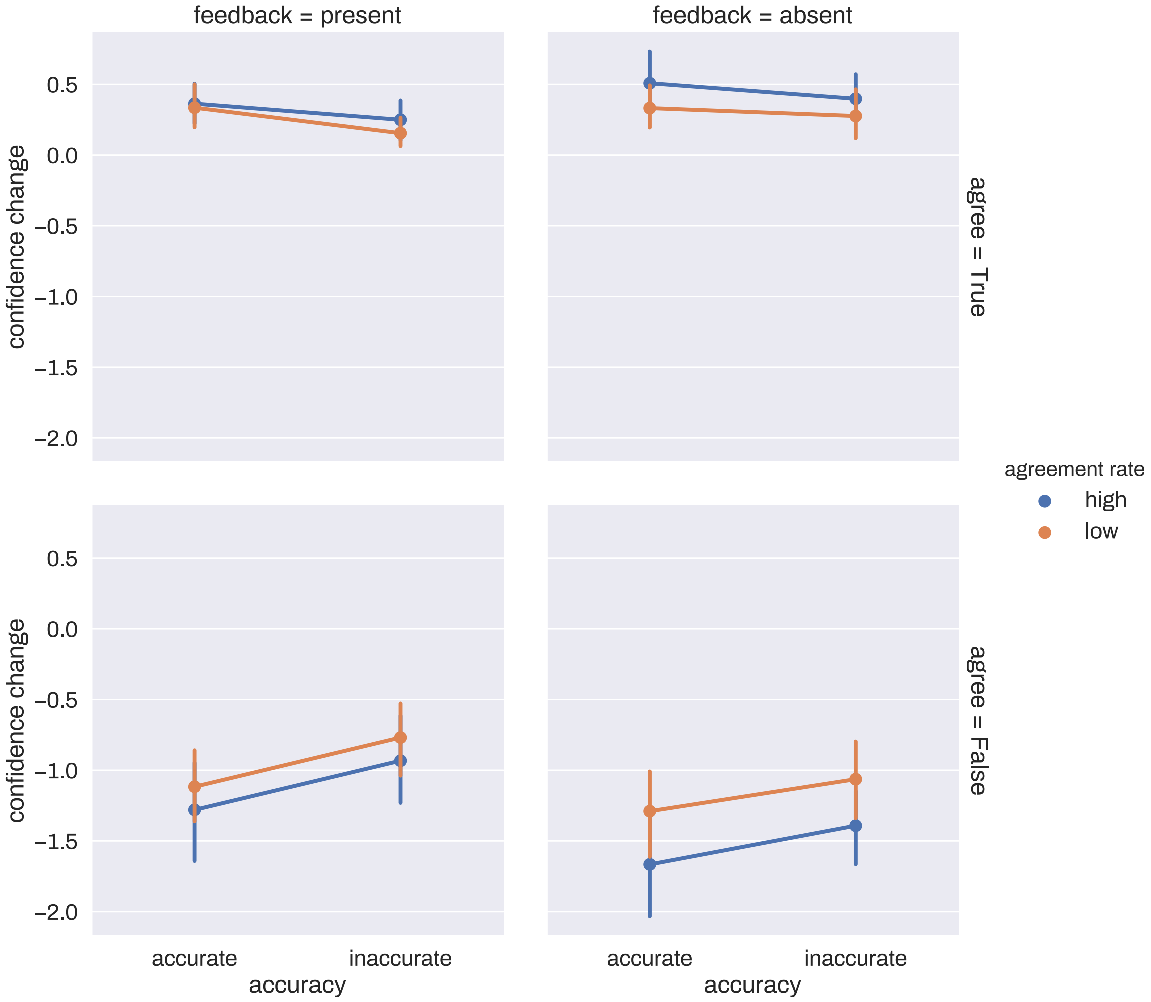

Figure S3: Confidence change in agreement and disagreement trials for different advisors. We observe an effect of both accuracy and agreement rates (see table S3) in both agreement and disagreement trials. This suggests that people accumulated a representation of others' competence irrespective, instead of simply discounting disagreeing evidence.

Effect	F(1,44)	p	ges
Agreement			
fb	0.9773923	0.328249136	1.718636e-02
acc	11.0912363	0.001762675 **	2.126074e-02
agr	11.6915961	0.001364394 **	1.796254e-02
fb:acc	0.8654221	0.357301932	1.692091e-03
fb:agr	2.0170753	0.162585940	3.145717e-03
acc:agr	0.0122558	0.912353006	1.609034e-05
fb:acc:agr	1.1344127	0.292646089	1.487152e-03
Disagreement			
fb	3.50257108	0.067931311 .	4.962705e-02
acc	13.31128416	0.000695565 ***	4.141194e-02
agr	12.24407362	0.001081278 **	3.128264e-02
fb:acc	0.35848098	0.552421231	1.162077e-03
fb:agr	1.64668779	0.206128216	4.324237e-03
acc:agr	0.03461100	0.853268206	6.699501e-05
fb:acc:agr	0.03595038	0.850489994	6.958741e-05

Table S3: We find the same main effects reported for influence reported in the main text. Columns from left to right: Effect, F statistic (numerator's degrees of freedom, denominator's degrees of freedom), p-value, generalized η_G^2 measure of effect size. Effect abbreviations: fb (feedback), acc (advisor accuracy rate), agr (advisor agreement rate). Significance values: . (< .1), * (< .05), ** (< .01), *** (< .001).

2.4 Experiment 3

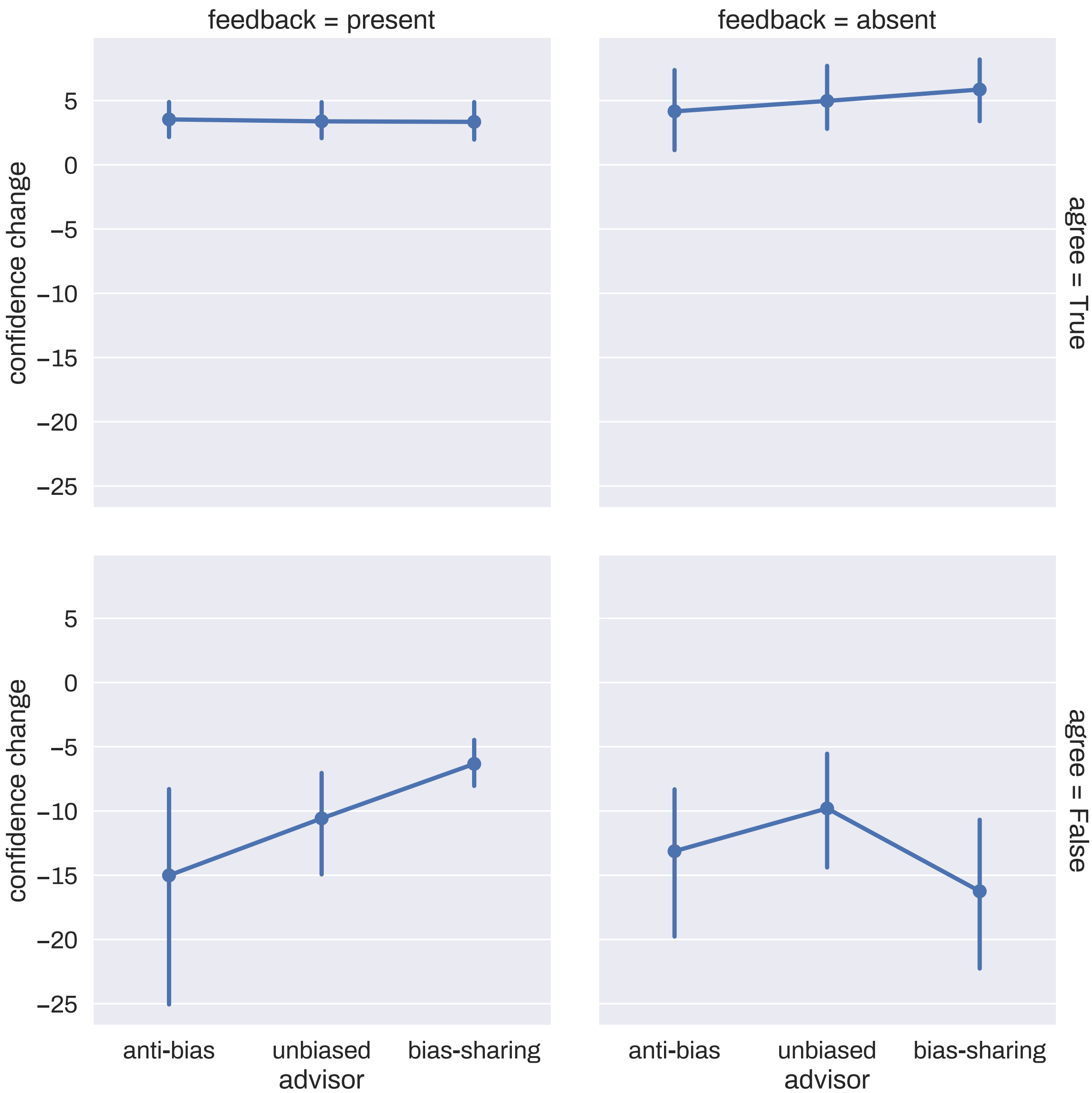

Figure S4: Confidence change in agreement and disagreement trials for different advisors. We observe an interaction between feedback and advisor type in both agreement (n.s.) and disagreement trials (Table S4). This suggests that people accumulated a representation of others' competence irrespective, instead of simply discounting disagreeing evidence.

Effect	DFn	DFd	F	p	ges
Agreement					
fb	1	46	1.3045963	0.2592860	0.022652313
adv	2	92	0.8914123	0.4135879	0.003529282
fb:adv	2	92	1.3948125	0.2530714	0.005511361
Disagreement					
fb	1	46	0.5992415	0.44282813	0.007701738
adv	2	92	1.6127341	0.20492582	0.013972922
fb:adv	2	92	4.2738736	0.01679238 *	0.036194829

Table S4: Experiment 3: confidence change broken down by agreement. In agreement trials, we observe the same numerical trend as reported in the main text, although these results are not significant. In disagreement trials on the contrary, the same interaction between feedback condition and advisor type is observed as the one reported in the main text, suggesting that disagreement trials might have been driving the main results. One possibility for this difference is the ceiling effect observed in agreement trials, which left more room for confidence change in the disagreement than in the agreement part. Columns from left to right: Effect, numerator's degrees of freedom, denominator's degrees of freedom, F statistic, p-value, generalized η^2_G measure of effect size. Effect abbreviations: fb (feedback), adv (advisor type). Significance values: . ($< .1$), * ($< .05$), ** ($< .01$), *** ($< .001$).

3 Three heuristics for advisor competence estimation

We used the models to explore how people would estimate advisor reliability if they were using three simple heuristic algorithms that make use of readily available information in the decision process: objective feedback, if available; the degree to which advisors agree with the participants' own initial judgments; and advisors' agreement with the participants' own judgments, scaled by the confidence with which those initial judgments are made. Notice that these models do not intend to be a faithful representation of the cognitive underpinnings of our human participants but a proof of concept showing that even in the absence of external feedback, confidence and agreement signals can be accumulated over time to form stable impressions of advisor's accuracy and reliability. Importantly, once these impressions are formed, they can inform a more flexible use of advice. For example, instead of simply down-weighting advice by confidence (as prior studies have shown), a stable representation of the advisor's underlying accuracy can be used to down-weight their advice also when their advice agrees with one's own current opinion. However, as thoroughly investigated in the main text, this strategy relies on the independence of one's own and one's advisor's judgments. If this independence is broken (e.g. if the self and the advisor are more likely to agree on incorrect choices) then this adaptive strategy backfires because it systematically overestimates the accuracy of highly agreeing individuals.

3.1 Model Description

Experiment 1 showed that people are sensitive to similar dimensions in the advice they receive both when objective feedback is available and when it was not provided. It is unclear what cues people are following to estimate their partners' reliability when feedback is taken away. Two simple explanations can be offered. The first one is that different advisors agreed differently often with participants and this in turn was taken as an indicator of good performance. In a binary choice task, if we assume that two people's judgments are independent, agreement rate between the two will scale linearly with the accuracy of each individual as long as performance is above chance. Thus, accumulating the number of agreement events over time for each individual advisor separately allows a subject to form a stable opinion about the other person's underlying accuracy.

A related, but subtler and potentially more powerful, strategy participants could have used is to accumulate over time the estimated probability of the advice being correct on a given trial. This quantity can be generated based on internal metacognitive signals of confidence, which provide a representation (albeit imperfect) of the uncertainty associated with a given perceptual judgment. In other words, given that confidence in a decision is a probabilistic representation of the correctness of that decision, it can also be used to estimate the likelihood that the advice received is correct or incorrect. Accumulating such evidence over time can help a decision maker to estimate the reliability underlying advice whenever more secure signals are not available.

To formalise such hypotheses we implemented a simple model that uses different pieces of information depending on different experimental conditions to estimate advisors' reliability. This simple model can then be compared with human observers to provide insight into the strategies they are using to evaluate advice reliability. Three different model variants are described below that account for the Feedback condition, the No-Feedback condition without metacognitive insight and the No-Feedback condition with metacognitive insight respectively.

3.2 Accuracy Model

When objective feedback is given to participants by the experimenter, the model can use it to infer the accuracy rate of its advisors. The accuracy of the advisor ($Acc = \{0, 1\}$) is the same as the accuracy of the subject in agreement trials, while is opposite in disagreement. By counting correct and error rates for each advisor separately, the model obtains a trial-by-trial estimation of the advisor's accuracy rate, θ, as the ratio between the number of advisor's correct trials and the total encounters with that advisor:

$$\theta^i = \frac{\alpha^i}{\alpha^i + \beta^i} \tag{5}$$

where α^i and β^i are the correct and error counts respectively, during the past trials with advisor i:

$$\alpha^i = \sum_{t=1}^{n} Acc_t \tag{6}$$

$$\beta^i = \sum_{t=1}^{n} 1 - Acc_t \tag{7}$$

$t = 1$ here represents the first encounter with advisor i while $t = n$ represents the last one. A slight complication in Experiment 1, however, is that advisors also provided a binary confidence judgment associated with the advice. A simple way for the model to make use of advisor's confidence is by treating it as a linear scaler of the advice received. We applied a set of arbitrary weights to the four possible advice scenarios, namely the advisor is (1) correct and confident, (2) correct but unsure, (3) incorrect and unsure and (4) incorrect but confident (Table S5). Although arbitrary, any set of weights that preserves the order of such events would result in similar final advisor preferences.

	Event Observed			
	Inaccurate Confident	Inaccurate Unsure	Accurate Unsure	Accurate Confident
Feedback	-1	-0.5	+0.5	+1
	Disagree Confident	Disagree Unsure	Agree Unsure	Agree Confident
No-Feedback	-1	-0.5	+0.5	+1

Table S5: Model weights (w) applied to different advice events observed in the Feedback and No-Feedback scenario.

Thus instead of simple accuracies, α and β in equations 6 and 7 can now be reformulated as:

$$\alpha^i = \sum_{t=1}^{n} .5 + .5 * w_t \tag{8}$$

$$\beta^i = \sum_{t=1}^{n} .5 - .5 * w_t \tag{9}$$

This set of equations results in values of 1, 0.75, 0.25 and 0 for the four events listed above respectively. Although these values could be simply summed to obtain α and β values, the unusual formulation of the equations 8 and 9 was preferred to be coherent with the equations describing the following models. They show how a simple model can take into account feedback, advice received and advisors' expressed confidence to track over time the objective reliability of its advisors.

3.3 Consensus Model

When feedback is removed from the participants, as in the No-Feedback conditions of our experiments, the model does not have access to the advisors' objective accuracy. It must then rely on different proxies for objective accuracy and integrate those instead over time. The first cue to underlying accuracy rate we considered is agreement rate. When two independent agents express judgments on a binary task, the agreement rate between the two linearly scales with the accuracy of each whenever the accuracy rate is higher than chance: $Agr = Acc_1 * Acc_2 + (1 - Acc_1) * (1 - Acc_2)$. We thus adapted the equations of the *Accuracy* model above to exploit this covariation. Instead of tracking the accuracy rates of its advisors, the *Consensus* model tracks their agreement rates with subjective judgments. Thus equations 8 and 9 can be used to estimate a θ value by now using as

w_t the scaled agreement observed on encounter t as described in Table S5. To take into account the fact that in Experiment 1 advisors expressed a binary confidence judgment themselves associated with the advice, we used the same linear weights applied to the *Accuracy* model also to scale agreement (Table S5). In other words this model perfectly conflates accuracy with agreement, assuming that whenever an advisor agrees with the subjective original judgment, the advisor must be correct. Although this clearly is a simplifying assumption, the model offers a useful proof of concept to understand what inferences an agent lacking metacognitive insight can make simply by using heuristics. It can thus provide a benchmark to quantify the information that is present in the advice received.

3.4 Confidence Model

A more nuanced strategy that could be employed to estimate advisors' reliability when feedback is not directly available is through use of internal metacognitive signals. Trial-level variability in subjective confidence is known to covary with objective accuracy in a perceptual task (Henmon, 1911) and it theoretically represents the estimated likelihood of having made a correct judgment and/or selected the correct response (Pouget, Drugowitsch, & Kepecs, 2016). Thus, instead of simply using agreement rates as a cue for accuracy rate, a model endowed with metacognitive insight could accumulate over time the subjective probability that an advisor expressed a correct judgment. A *Confidence* model was created under the assumption that the trial-by-trial subjective reports of confidence are directly related to the true underlying estimated probabilities of having chosen the correct answer. On agreement trials the model estimates the probability of the advice being correct as the subjective probability of a correct answer. Conversely on disagreement trials the model estimates the probability of the advice being correct as the probability of having itself made an error. In other words trial-level agreement ($Agr = \{0, 1\}$) is scaled by trial-confidence expressed as a probability over outcomes (correct vs. incorrect response). Thus equations 8 and 9 above become according to this model:

$$\alpha^i = \sum_{t=1}^{n} .5 + (p_t(corr) - .5) * w_t \tag{10}$$

$$\beta^i = \sum_{t=1}^{n} .5 - (p_t(corr) - .5) * w_t \tag{11}$$

where w_t represents the scaled trial-level agreement as described in Table S5 and $p(corr)$ represents pre-advice confidence. As described below, rather than taking participants' confidence as a pure index of subjective $p(corr)$, we transformed the value to (1) reduce inter-subjects variability and (2) increase scale sensitivity. Regardless, the crucial point is that this model capitalises on the fact that being in agreement or disagreement with an advisor is more informative when the model is itself confident that it gave a correct answer than when it is more likely to have made a mistake.

Experiments 2-3. In Experiments 2 and 3, advisors did not express a level of confidence with their judgments. This allowed to simplify the above equations describing the three models. The *Accuracy* model could be simplified using equations 6 and 7 to compute α and β for each advisor instead of equations 8 and 9. Similarly, the *Consensus* model now computes α and β values for each advisor i separately as:

$$\alpha_i = \sum_{t=1}^{n} .5 + .5 * Agr_t \tag{12}$$

$$\beta_i = \sum_{t=1}^{n} .5 - .5 * Agr_t \tag{13}$$

where Agr_t is the partner's consensus ($Agr = \{-1, 1\}$) observed on encounter t. Finally, the simplified *Confidence* model computes α and β values as:

$$\alpha = \sum_{t=1}^{n} .5 + (p(corr) - .5) * Agr_t \tag{14}$$

$$\beta = \sum_{t=1}^{n} .5 - (p(corr) - .5) * Agr_t \tag{15}$$

where $p(corr)$ is the pre-advice confidence expressed in probability scale as described in equation 18.

3.5 Bayesian update

All model variants can use the current estimated advisor's reliability θ to appropriately update the pre-advice probability of having selected the correct answer $p(corr)$ into a normative posterior, based on the binary advice A received (agree vs. disagree):

$$p(corr|A^i) = \frac{p(corr)p(A^i|corr)}{p(corr)p(A^i|corr) + p(err)p(A^i|err)} \tag{16}$$

where $p(err)$ is the subjective probability of making a mistake on the current trial and $p(A^i|corr)$ is the probability that advisor i agrees or disagrees given that the participant's choice is correct. Prior probability $p(corr)$ is estimated from a simple linear transformation of the pre-advice trial-level confidence data obtained from the participants after appropriate pre-processing. Pre-processing consisted in a parameter-free transformation that (a) brings all subjective confidence distributions on to a similar scale thus reducing the inter-subject variability and (b) expands the centre of the original subjective confidence distributions so to increase the informativeness of the average trial. This operation was inspired by recent models of adaptive information gain control (Cheadle et al., 2014). According to these proposals, the brain adapts the gain of neuronal firing to the range of information available over different time scales and cognitive domains (Carandini & Heeger, 2011; Cheadle et al., 2014). Here it serves the purpose of increasing the discriminability or information gain of different trials so that trials that are close together on confidence scale gets pulled apart on to a probability scale. The transformation uses parameters obtained from the data:

$$\hat{C}_{pre} = N * normcdf(C_{pre}) \tag{17}$$

where $normcdf(C)$ is the normal cumulative density function of the pre-advice confidence C ratings distribution, and N is the number of confidence ratings available on each interval of the scale (in Experiments 1,2: N = 5; in Experiment 3: N = 50). This simple transformation has the property of translating a normal distribution into a uniform distribution in the range $[0, N]$. Notice that this transformation does not affect the ranking of confidence judgments but only their spacing along a probability scale. After pre-processing, confidence ratings were translated into a probability scale with the linear transformation:

$$p(corr) = 0.5 + (0.1 - \epsilon) * \hat{C}_{pre} \tag{18}$$

where ϵ is a small jitter ($\epsilon = .002$) introduced to avoid maximum confidence ratings being turned into probability of one and zero, which would in turn cause inconsistencies within the Bayesian formula (e.g., no confidence change regardless of advice reliability). Thus $p(corr)$ represents trial-level confidence on a probability scale, which can be interpreted as the probability that the participant assigns to having given a correct answer on a given trial. From $p(corr)$ we can also derive the subjective probability that a given trial will end up in an error: $p(err) = 1 - p(corr)$.

To estimate the likelihood term $p(A^i|corr)$ in equation 16 we applied a simple heuristic that uses the reliability θ of a given advisor:

$$p(A^i|corr) = \theta^A * (1 - \theta)^{1-A} \tag{19}$$

The equation above simply states that the probability of observing advisor i's agreement ($A^i = 1$) when the participant is correct is equal to the accuracy rate of the advisor itself, assuming advisor's and participant's judgments are independent. Conversely, the probability of observing disagreement ($A^i = 0$) on the same trials is the advisor's error rate. In other words, the probability of agreement in trials when the participant is correct is the probability that the advisor too is correct. Similarly, the probability of disagreement in trials when the participant is correct is equal to the probability that the advisor is wrong.

3.6 Model results

Trial-by-trial agreement, reported confidence and objective feedback from the experimental data of the 46 participants in Experiment 1 were used to estimate rated competence and influence that the our three heuristic model variants would show with each advisor if they had experienced the corresponding advice profiles of the four virtual advisors. Separate model runs simulated the evolution of the accuracy estimate (theta parameter) according to the three learning rules described above: the *Accuracy* model that learns based on trial-by-trial feedback, which by hypothesis should capture patterns of rated competence and influence from participants the Feedback condition, and the *Consensus* and *Confidence* models, which provide distinct computational accounts of the evolution of rated competence and influence in the No Feedback condition—whether it depends purely on rates of agreement, or whether agreement is weighted according to participants' confidence in their own initial judgments.

Experiment 1 The three model variants were applied to the actual series of each participant's decisions and (where appropriate) their associated confidence, and the advice they received and (where appropriate) its accuracy, in Experiment 1. The aim of the following analyses was to verify how the pattern of final model's trust (Θ) in each advisor differed when different pieces of information were used to compute it. The models are not intended as a mechanistic description of participants' behaviour, but rather aim to explore how simple accuracy estimation rules lead to differentiated patterns of trust across advisors according to the type of information used to update estimates of advisor accuracy. For this analysis, data from the Feedback and No-Feedback groups were pooled together to increase statistical power, as the presence of feedback did not affect the variables that model's variants were based on, namely advisors' accuracy, agreement rates and participant's pre-advice confidence ratings respectively. Our analysis focuses on the resulting values for each model variant across simulated participants, as a direct measure of the model's belief about advisor accuracy.

The three model variants' final Θ values were analyzed using a 2x2 repeated measures ANOVA with factors of advisor Accuracy (high vs low) and Calibration (high vs. low). Scaling factors were applied to agreement to take into account the fact that in this experiment advisors provided a confidence judgment with their advice. The Accuracy variant is, in this experiment, fully pre-determined by the advisors' set accuracy rates and thus no statistical analysis was run due to the absence of variability across participants. We plot however its trust value for visual reference as it shows that when the model is provided with information about the objective performance of the participant (and thus of the advisors), it is able to distinguish advisors both in terms of their Accuracy rate and their confidence Calibration. The difference between calibrated and uncalibrated advisors is larger for inaccurate than accurate advisors, due to the weights used to convey advice confidence (Figure S5).

In the absence of objective feedback, both the Consensus variant—which estimates advisors' reliability by assuming that advisors are correct whenever they agree with the participant's own first decision, and wrong otherwise—and the Confidence variant—which uses agreement as a proxy for feedback like the Consensus variant, but scales them by pre-advice confidence—show greater trust for Accurate ($F(1, 45) > 165.84, p < .001, \eta_G^2 = .44$) and Calibrated ($F(1, 45) > 32.70, p < .001, \eta_G^2 = .13$) advisors compared to inaccurate or uncalibrated ones. Neither variant showed a significant interaction between the two factors ($F(1, 45) < 2.65, p > .11, \eta_G^2 = .01$). Notice that advisors were not constrained to agree with the participant a pre-determined number of times, thus explaining the variability observed across participants according to the specific sequence of decisions they made and advice they received. Taken together, these modeling results show that simple computations of advisor reliability perform well at this task even when trial-level feedback is absent, effectively capturing key patterns of trust observed in the human data across feedback conditions. For the task used in Experiment 1, the three variants do not make contradictory predictions on which advisors should be trusted. In particular the two No-Feedback variants (that base trust on simple agreement, or agreement weighted by confidence) cannot be disentangled using the data collected from this experiment, with both showing sensitivity to both accuracy and calibration of an advisor.

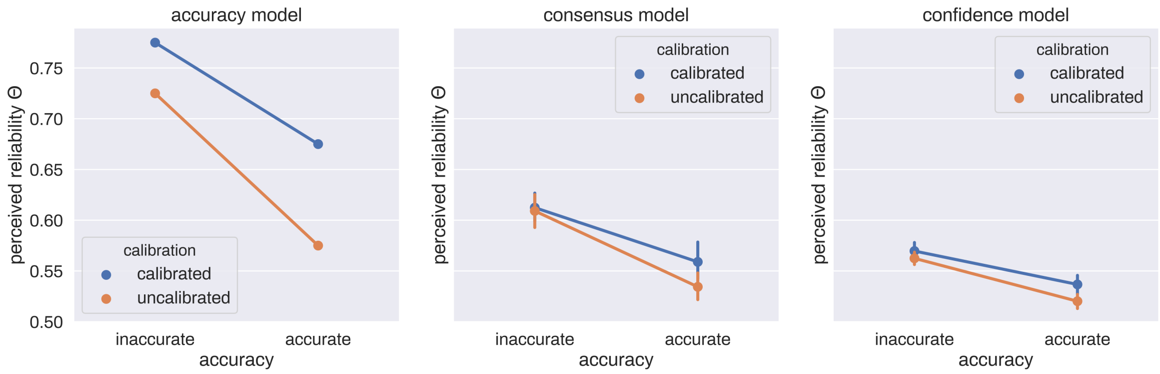

Figure S5: Experiment 1 - Heuristic models

Experiment 2 The models described above were applied to data from Experiment 2, to understand whether the Consensus and Confidence model variants behaved differently in scenarios where advice accuracy and advice agreement rate are dissociated. In this experiment, advisors did not express a confidence judgment about their opinions. Thus, all model variants could be simplified by not taking into account advice confidence. Trial-by-trial pre-advice confidence and advice were input to each of the three model variants and resulting Θ-values for each advisor were compared (Figure S6). Both the Accuracy and the Confidence models' Θ values showed a significant effect of Accuracy ($F(1, 45) > 8.85, p < .005, \eta^2_G > .05$), while the Consensus model only showed a non significant marginal effect ($F(1, 45) = 3.22, p = .07, \eta^2_G = .02$). Both the Confidence and Consensus models show a significant effect of Agreement ($F(1, 45) > 434.7, p < .001, \eta^2_G > .71$), but no reliable interaction between the two factors ($F(1, 45) < 2.04, p > .15, \eta^2_G < .007$).

Not surprisingly, when provided with objective feedback on trial-by-trial performance, a simple model of reliability estimation (Accuracy variant) distinguished advisors based on their accuracy but not their agreement profile. More surprisingly, a model without access to feedback but endowed with metacognitive insight (Confidence variant) was also able to discriminate between equally agreeing but differently accurate partners. As shown in Table 2 (main text), the accurate agreeing advisor tends to agree more often than the inaccurate agreeing advisor when the participant is objectively correct (6.5 times out of 7 against 5.5 times out of 7) and less often when the participant is objectively wrong (1.5 times out of 3 against 2.5 times out of 3). Trials when participants' initial judgment is correct are usually associated with greater confidence ratings (Fleming et al., 2014; Henmon, 1911; Koriat, 2012), thus a strategy of reliability estimation relying on confidence can exploit this covariation to detect differences in accuracy, notwithstanding equal agreement rates.

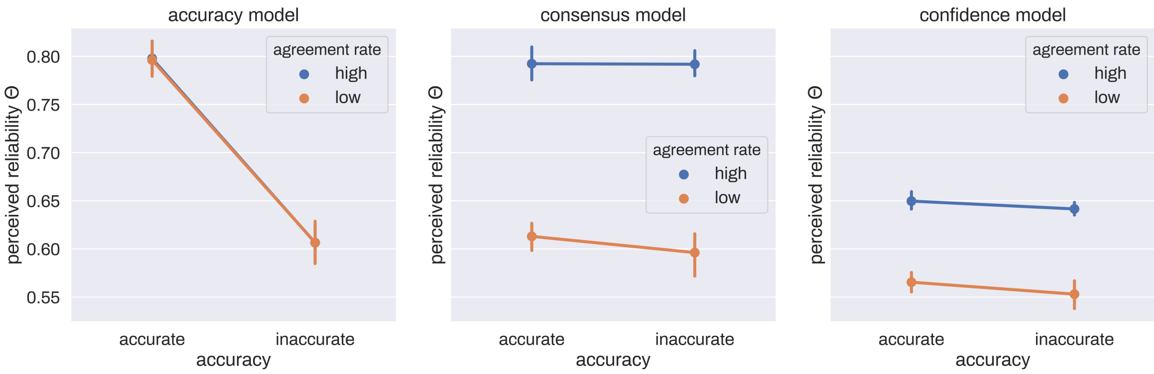

Figure S6: Experiment 2 - Heuristic models

Experiment 3 The following simulations and analyses explored the differing patterns of trust across advisors predicted by simple models of estimating advisor reliability. Simulations are also useful in this experiment as a check that our designs were controlled as intended (e.g., for agreement rates across advisors) even though we had less precise control over conditions because counterbalancing depended on an evolving estimate of participants' confidence distributions. Figure S7 shows the pattern of results (modeled -values) that the three model variants produce.

Both when the model has access to trial-by-trial feedback (Accuracy variant), and when it only has access to past agreement (Consensus variant), no significant effect of Advisor is observed ($F(2, 94) < 1.70, p > .18, \eta^2_G < .02$), nor is there a difference between the bias-sharing and the anti-bias advisor. These patterns are expected because the three advisors were matched for accuracy and agreement rates by design in this experiment. On the contrary, a Confidence variant which uses metacognitive information and past agreement (but lacked access to trial-level feedback) showed a significant effect of Advisor ($F(2, 94) = 7.95, p < .001, \eta^2_G = .10$). Specifically, simulated trust was higher for the bias-sharing advisor than the anti-bias advisor ($t(47) = 3.54, p = .001, d = .74$), and higher for the unbiased advisor than the anti-bias advisor ($t(47) = 2.99, p = .004, d = .57$). Simulated trust was higher for the unbiased than the bias-sharing advisor, but this difference was not reliable ($t(47) = 1.21, p = .22, d = .25$). These findings indicate that by accessing metacognitive signals (as provided in the model by participants' confidence ratings) the model was able to discriminate among different advisors. This model correctly predicts greatest levels of trust in a bias-sharing advisor, but also predicts lowest levels of trust in anti-bias advisors, whereas our experimental participants expressed (numerically) lowest levels of trust in unbiased advisors.

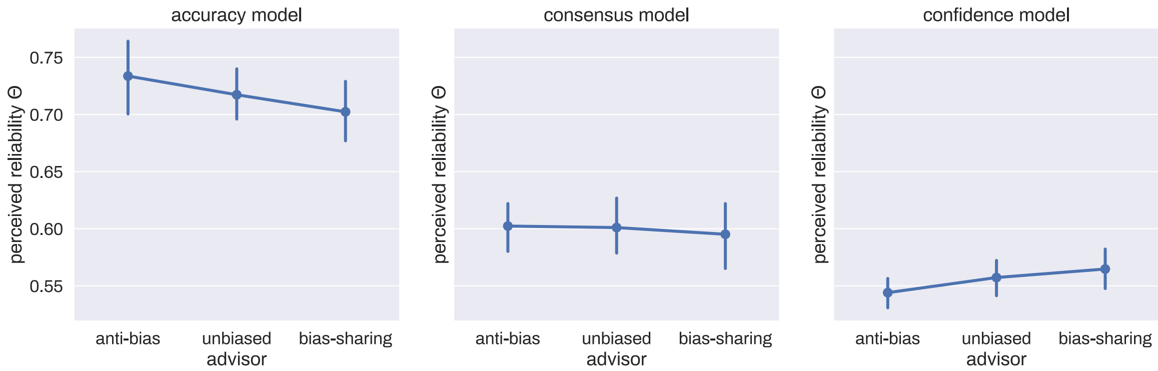

Figure S7: Experiment 3 - Heuristic models

4 Post-advice confidence correlations.

Our main analyses for each experiment focused on qualitative predictions arising from different strategies for inferring advisor reliability. Collectively, the results are consistent with the hypothesis that people use their internal sense of confidence in making these inferences—showing sensitivity to advisor accuracy in the absence of objective feedback, even when advisors are matched for agreement rate, and developing differing patterns of trust when advisor agreement rates vary with their own expressed decision confidence. Our final analysis of the empirical data attempted a more quantitative comparison of model predictions, specifically focusing on whether the *Consensus* or *Confidence* models better predicted post-advice confidence ratings across trials for the participants in the No Feedback conditions of Experiments 1-3.

For this analysis we used Bayes rule to infer the trial-by-trial post-advice confidence ratings that each variant would express given a participant's expressed pre-advice confidence and advisor agreement (as defined above). The within-participant correlation between participants' post-advice confidence and model's post-advice confidence was computed for each experiment, for No-Feedback groups only. Second-order statistics were performed to test, across experiments, which variant was more strongly correlated with the human data. A 2x2 ANOVA on correlation coefficients with Model (*Consensus* vs. *Confidence*) and Experiment as factors showed that the *Confidence* variant was significantly more correlated with participants' responses than the *Consensus* variant ($F(1, 22) = 8.18, p = 0.009, \eta^2_G = 0.0049$). No significant effect of experiment nor interaction between the two were found ($F(2, 44) < 1.3, p > .25$), suggesting that, across experiments, the *Confidence* model's post-advice confidence more strongly covaried with participants' true post-advice responses. As a check for the soundness of this model comparison method, the same 2x2 ANOVA was run on the correlation coefficients between participants' post-advice confidence and the model's post-advice confidence predictions, after randomly shuffling trials within each participant. This operation should ensure that any advantage of the *Confidence* variant over the *Consensus* variant is not due to unspecific factors (like being overall more conservative in updating confidence), but rather to trial-level variability. After reshuffling, the *Consensus* and *Confidence* variants were not significantly different from each other ($F(1, 22 = 1.95, p = 0.17, \eta^2_G = 9.54e - 04)$, corroborating our conclusions.

5 Agent-based simulation

5.1 Model description

An agent-based model was programmed using NetLogo (Wilensky, 1999) and is available at `https://github.com/chri4354/trust_formation_without_feedback`. The model was initialised as a fully connected directed network of N agents. A directed edge from agent i to agent j represents the trust $\theta_{i,j}$ that i has in j's opinions. We simulate agents on a lattice network performing repeated binary A/B decisions, receiving advice from other agents, inferring their reliability and updating their own initial decisions. We let the simulation run for a 1000 steps. A signal s with strength S was drawn from a uniform distribution between $-\frac{S}{2}$ and $+\frac{S}{2}$. This represents the decision quantity to estimate (e.g., difference in dots or true state of the world). The task of each agent was to determine if s was positive (event A) or negative (event B). Each agent estimated the posterior probability of A given the perceptual information generated by s as follows:

$$p'(A) = p(A|E_p) = \frac{p(A)E_p}{p(A)E_p + p(\bar{A})\bar{E_p}} \tag{20}$$

$$E_p = L(s + \mathcal{N}(0, \sigma)) \tag{21}$$

where $p(A)$ is the prior probability of observing As before seeing any stimulus, L is a logistic sigmoid mapping from sensory evidence to probability; E_p is the perceptual evidence resulting from such mapping and $\mathcal{N}$ is independent individual perceptual Gaussian noise with mean 0 and standard deviation σ. Bars represent complement probability. Each agent's perceptual noise (and thus accuracy) was manipulated by varying the noise parameter σ. Each agent's bias was manipulated by varying the initial value $p(A)$. Agents' confidence was represented as the distance from the uncertainty point 0.50:

$$C = .50 + |p'(A) - .50| \tag{22}$$

Trust, represented by the network's edges, was initialized to 0.50 for every agent and updated after social interaction. After making a judgment, agents selected one other agent to interact with either at random (random sampling) or proportionally to their trust (biased sampling). Agents then updated their initial judgment $p'(A)$ as follows:

$$\hat{p}(A) = p(A|E_s) = \frac{p'(A)E_s}{p'(A)E_s + p'(\bar{A})\bar{E_s}} \tag{23}$$

where E_s represents social evidence and is obtained from the advisor's judgment either by taking the advisor's raw judgment $p'(A)$ (without advice discounting) or by discounting the advisor's judgment proportionally to the agent's trust in the advisor (with advice discounting). Advice discounting consisted in a linear regression toward the uncertainty point 0.50 using the following equation:

$$E'_s = 0.5 + (\theta * (p'(A) - 0.5)) \tag{24}$$

The above equation regresses any confidence judgment $p'(A)$ toward the uncertainty point 0.50 proportionally to trust. A trust level of 1 would leave the advisor's judgment $p'(A)$ untouched, while a trust level of 0 would make any advisor's judgment equal to 0.50 and thus entirely uninformative. After updating their judgments, each agent i updated its trust judgments (i.e., outward edges θ_i) based on the available information about other agents. If feedback is available, the agent updates its current trust in agent j by virtue of a delta rule in the form:

$$\theta_{i,j}^{t+1} = \theta_{i,j}^{t} + \alpha(F_j - \theta_{i,j}^{t}) \tag{25}$$

where F_j is the accuracy of agent j and α is a learning rate set to 0.1. If feedback is not available on the contrary, the agent replaces F with $\hat{F}$, or the *estimated* partner's accuracy. $\hat{F}$ was calculated using the agreement or agreement-in-confidence heuristics described above. In our simulations, we assessed the effect of feedback availability as it varied parametrically, from being available after every decision (i.e., p-feedback = 1.0), available after only some decisions (i.e., 0 < p-feedback < 1), or never available (i.e., p-feedback = 0), rather than the simpler case of feedback presence/absence that we studied experimentally above.

The emergence of trust patterns when using agreement-based heuristics can be expected to track true accuracy when judgments are independent but generate clustering of populations when judgment correlations emerge within such populations. We defined a network's clustering coefficient as the ratio between average trust toward agents who initially share the same bias (in-group trust) and total average trust: $\bar{\theta}_{in-group}/(\bar{\theta}_{in-group} + \bar{\theta}_{out-group})$. A ratio of 0.5 represents no preference (i.e., no difference in trust) toward agents sharing the same initial bias, while a ratio greater than 0.5 represents a preference toward agents sharing the same initial biases. We test how network clustering is shaped by the presence or absence of objective feedback and show that bias-specific segregation arises only when feedback is rarely available.

Finally, once bias-specific segregation is established, we ask whether such clustering remains stable. In particular, after 500 iterations we allow agents to dynamically change their original bias as a function of experience. For example, it is known that the bias observed in people performing binary judgments is influenced by their recent history of decisions (Akaishi, Umeda, Nagase, & Sakai, 2014; Zylberberg, Wolpert, & Shadlen, 2018). In the present context, if an agent systematically reports "A" but receives negative feedback, they should reduce their bias by decreasing their prior probability $p(A)$. Similarly, when feedback is absent, an agent who systematically reports "A" but finds themselves, after interacting with other agents, believing that Bs are more frequent than expected, should reduce their bias towards As. Conversely, bias should get stronger if the social contexts reinforces it (although see (Bail et al., 2018)). We modelled bias update with a delta rule:

$$p(A)^{t+1} = p(A)^t + \alpha(I^t - p(A)^t) \tag{26}$$

where I^t is an indicator variable that represents the final belief in the event A. When objective feedback is available, I takes the value of 1 if an event A occurred and 0 otherwise. When feedback is not available, I is set to the discrete or continuous final subjective belief in the event A. In the following section, we show the results obtained when a discrete final belief is used in the absence of feedback:

$$I = \begin{cases} 1, & \text{if } \hat{p}(A) \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Similar results were obtained setting I to the continuous belief $\hat{p}(A)$.

The following figures supplement figures in the main text.

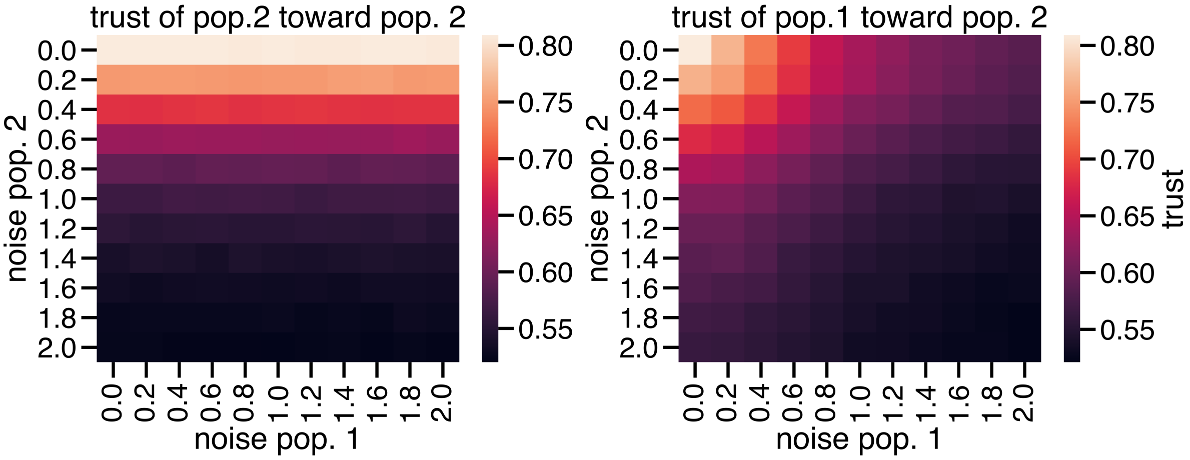

Figure S8: Trust of each subpopulation toward Population 2. Left panel: Trust of Population 2 towards Population 2 is inversely proportional to the noise of Population 2 agents (y-axis), but are (unsurprisingly) unaffected by the noise of Population 1 (x-axis). Right panel: Trust of Population 1 toward Population 2 is affected by both the noise of Population 2 and the noise of Population 1. Although the former correctly tracks Population 2's true underlying reliability, the latter reflects an interaction between the judge's characteristics and the advisor's characteristics, cf Kruger and Dunning (1999).

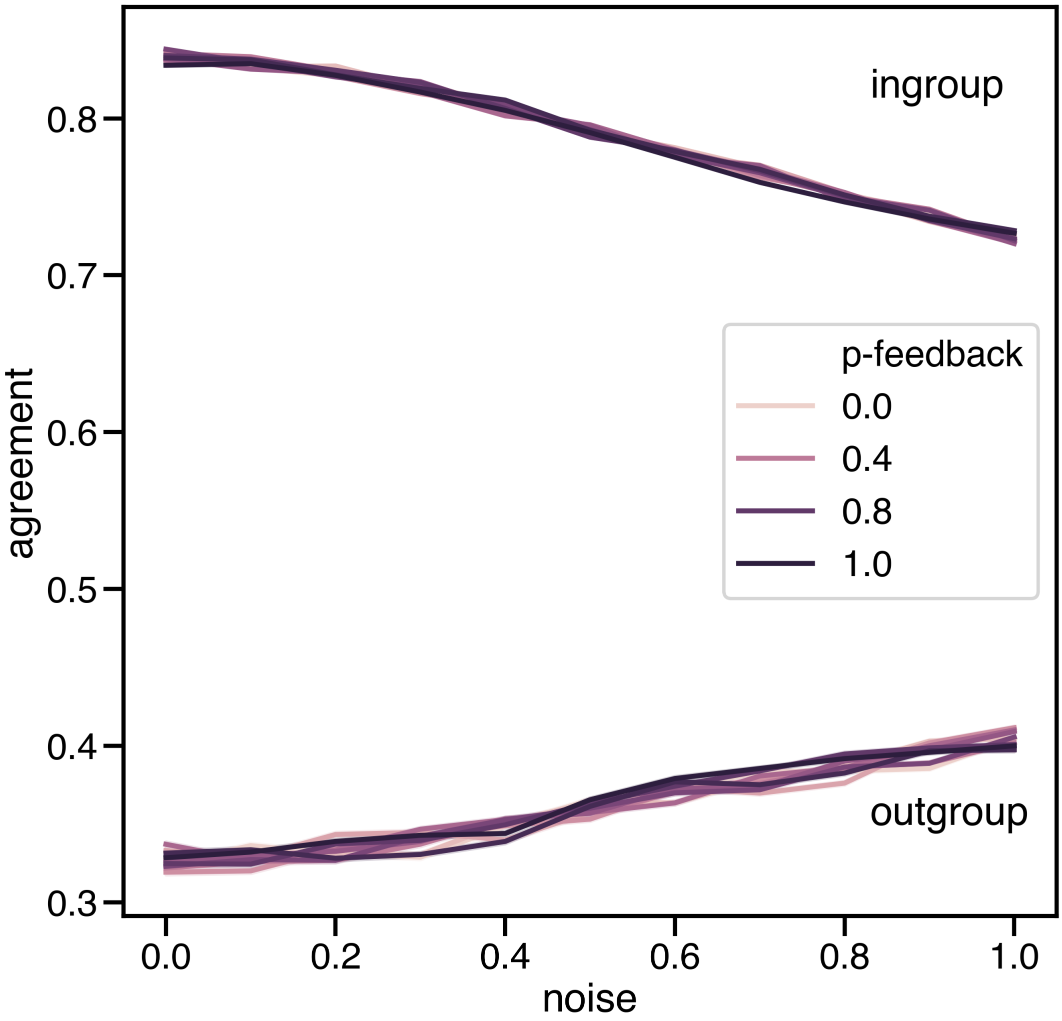

Figure S9: Average agreement rate as a function of probability of feedback and noise. Agreement with ingroup appears to decrease as a function of increasing noise, while agreement with outgroup tends to increase as a function of noise. On the contrary, feedback availability does not affect agreement rates.

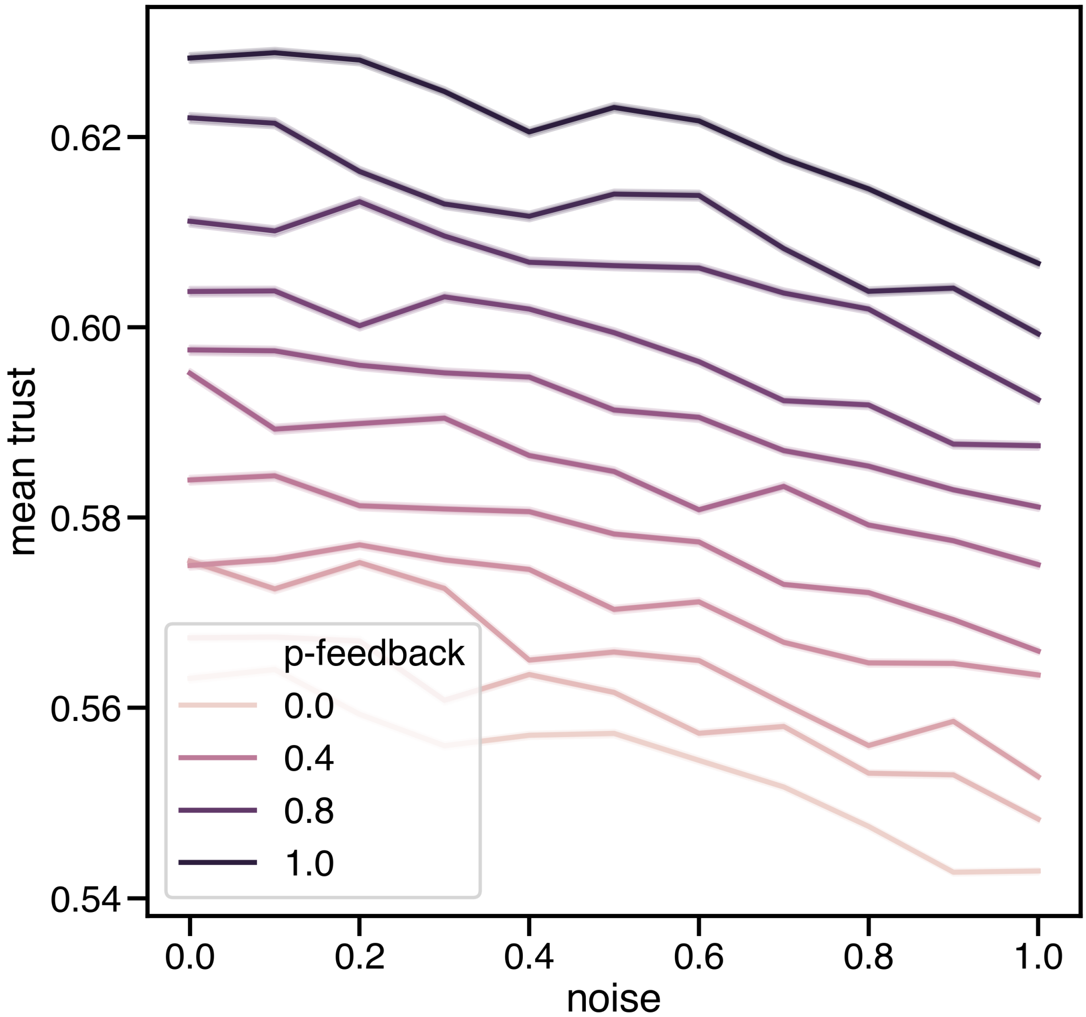

Figure S10: Average trust as a function of probability of feedback and noise. Trust appears to decrease as noise increases and feedback availability decreases.

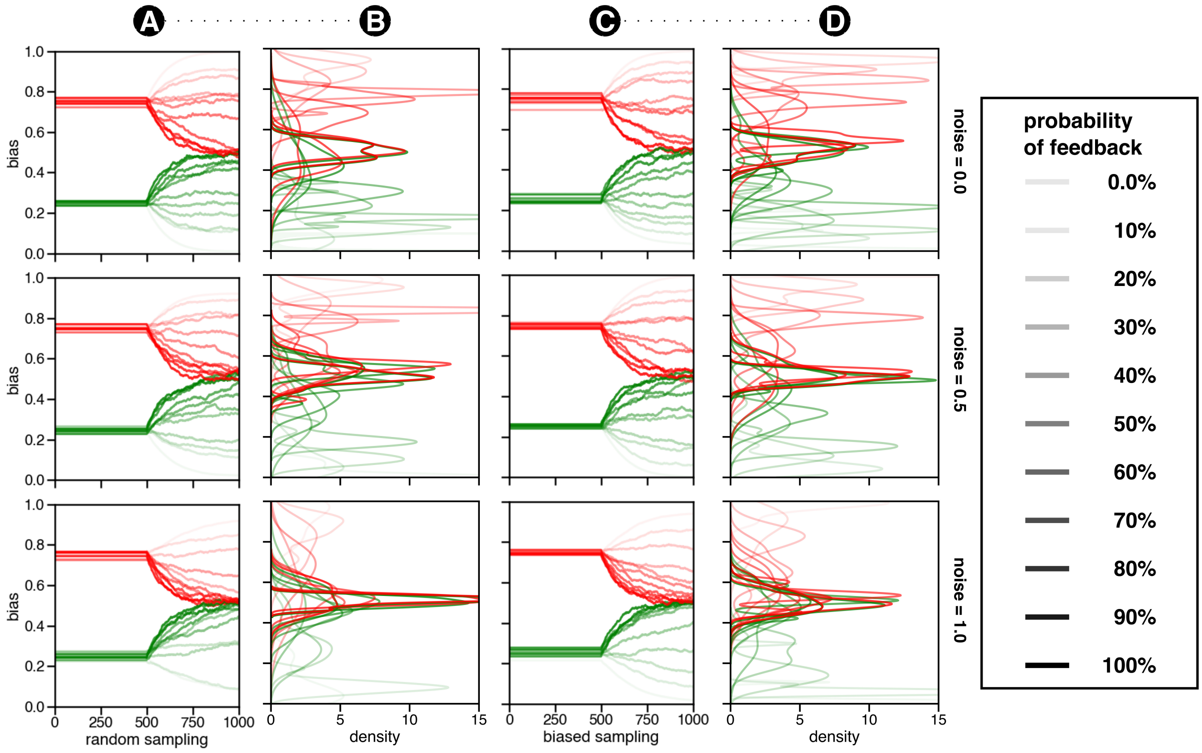

Figure S11: Bias distribution and evolution in simulations where agents discount advice proportionally to trust in the advisor.

References

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014, 1). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, *81*(1), 195–206. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/24333055` doi: 10.1016/j.neuron.2013.10.018

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., ... Volfovsky, A. (2018, 9). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. Retrieved from `http://www.pnas.org/lookup/doi/10.1073/pnas.1804840115` doi: 10.1073/pnas.1804840115

Carandini, M., & Heeger, D. J. (2011, 11). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. Retrieved from `http://www.nature.com/doifinder/10.1038/nrn3136` doi: 10.1038/nrn3136

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castañón, S., & Summerfield, C. (2014, 3). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–41. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/24656259` doi: 10.1016/j.neuron.2014.01.020

Fleming, S. M., & Lau, H. C. (2014, 7). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. Retrieved from `http://www.frontiersin.org/Human_Neuroscience/10.3389/fnhum.2014.00443/abstract` doi: 10.3389/fnhum.2014.00443

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2014). Action-Specific Disruption of Perceptual Confidence. *Psychological science*. doi: 10.1177/

0956797614557697

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186–201. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/h0074579 doi: 10.1037/h0074579

Koriat, A. (2012, 4). When are two heads better than one and why? *Science (New York, N.Y.)*, *336*(6079), 360–2. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1216549http://www.ncbi.nlm.nih.gov/pubmed/22517862 doi: 10.1126/science.1216549

Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of personality and social psychology*, *77*(6), 1121–1134.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016, 2). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. Retrieved from http://www.nature.com/doifinder/10.1038/nn.4240 doi: 10.1038/nn.4240

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... Nelson, C. (2009, 8). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, *168*(3), 242–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3474329&tool=pmcentrez&rendertype=abstract doi: 10.1016/j.psychres.2008.05.006

Wilensky, U. (1999). *NetLogo.* Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018, 9). Counterfactual Reasoning Underlies the Learning of Priors in Decision Making. *Neuron*, *99*(5), 1083–1097. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0896627318306330 doi: 10.1016/j.neuron.2018.07.035