AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2020 American Psychological Association
ISSN: 0096-3445

2021, Vol. 150, No. 3, 507–526
http://dx.doi.org/10.1037/xge0000960

The Role of Decision Confidence in Advice-Taking and Trust Formation

Niccolò Pescetelli
University of Oxford and Max Planck Institute for Human Development, Berlin, Germany

Nicholas Yeung
University of Oxford

In a world where ideas flow freely across multiple platforms, people must often rely on others' advice and opinions without an objective standard to judge whether this information is accurate. The present study explores the hypothesis that an individual's internal decision confidence can be used as a signal to learn the accuracy of others' advice, even in the absence of feedback. According to this "agreement-in-confidence" hypothesis, people can learn about an advisor's accuracy across multiple interactions according to whether the advice offered agrees with their own initial opinions, weighted by the confidence with which these initial opinions are held. We test this hypothesis using a judge-advisor system paradigm to precisely manipulate the profiles of virtual advisors in a perceptual decision-making task. We find that when advisors' and participants' judgments are independent, people can correctly learn advisors' features, like their accuracy and calibration, whether or not objective feedback is available. However, when their judgments (and thus errors) are correlated—as is the case in many real social contexts—predictable distortions in trust can be observed between feedback and feedback-free scenarios. Using agent-based simulations, we explore implications of these individual-level heuristics for network-level patterns of trust and belief formation.

Keywords: metacognition, confidence, advice-taking, trust, decision-making

Supplemental materials: http://dx.doi.org/10.1037/xge0000960.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

We rely on advice in many everyday contexts, from finance and politics to education and health, but often lack immediate feedback or other objective standards with which to judge the accuracy of that advice. Yet in these contexts we must learn to distinguish good from bad advisors and consequently who to listen to. How people do this, and how reliably they do so, are open questions that have received relatively little systematic study to date (see (Weiss & Shanteau, 2003) for exceptions). The present research addresses this issue.

Although trust is a multidimensional construct, in the current work we are interested in a specific aspect of trust pertaining to the accuracy and competence of the trustee (Mayer, Davis, & Schoorman, 1995). In relation to this, we ask three related questions. First, we ask whether people can learn their advisors' competence through experience even in domains where feedback is unreliable, costly, or completely absent, and in the absence of contextual cues (e.g., reputation). Second, we ask what heuristics people use to form competence representations in these contexts, and under what circumstances these heuristics are useful versus maladaptive. Third, we explore how heuristics used by individual decision makers can drive emerging patterns of trust and influence in larger groups. We find that people can discern the usefulness of advice without the benefit of external feedback, and that simple mechanisms can explain the observed behavior.

When available, external feedback can guide learning about others through reinforcement learning mechanisms that have been thoroughly described in previous research (Behrens, Hunt, Woolrich, & Rushworth, 2008; Guggenmos, Wilbertz, Hebart, & Sterzer, 2016; Sutton & Barto, 1998). But a similar kind of social inference is likely to be necessary even when such objective external feedback is not readily available. Our hypothesis is that a solution to this seemingly computationally intractable problem lies in the use of two readily available pieces of evidence in social decision making contexts—namely, an advisor's agreement rate with one's own beliefs and one's own internal confidence in those beliefs: If for a given decision we are certain we are correct, then we can equally be certain that anyone who disagrees with us is wrong, and accordingly down-weight their opinion in the future. If on the contrary we make a choice with less confidence, we should still down-weight our trust in that advisor in future interactions, but now to a lesser extent. Thus, people can overcome the absence of objective feedback by accumulating over time the trial-by-trial covariation between their internal decision confidence and the actual state of the environment (Guggenmos & Sterzer, 2017; Guggenmos et al., 2016; Pescetelli, Rees, & Bahrami, 2016). We call this strategy the *agreement-in-confidence* heuristic.

Our work builds on previous research into group decisions and information integration processes (Bonaccio & Dalal, 2006; Rader,

This article was published Online First October 1, 2020.

Niccolò Pescetelli, Department of Experimental Psychology, University of Oxford, and Max Planck Institute for Human Development, Berlin, Germany; Nicholas Yeung, Department of Experimental Psychology, University of Oxford.

Correspondence concerning this article should be addressed to Niccolò Pescetelli, Max Planck Institute for Human Development, 94 Lentzeallee, Berlin 14195, Germany. E-mail: pescetelli@mpib-berlin.mpg.de

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

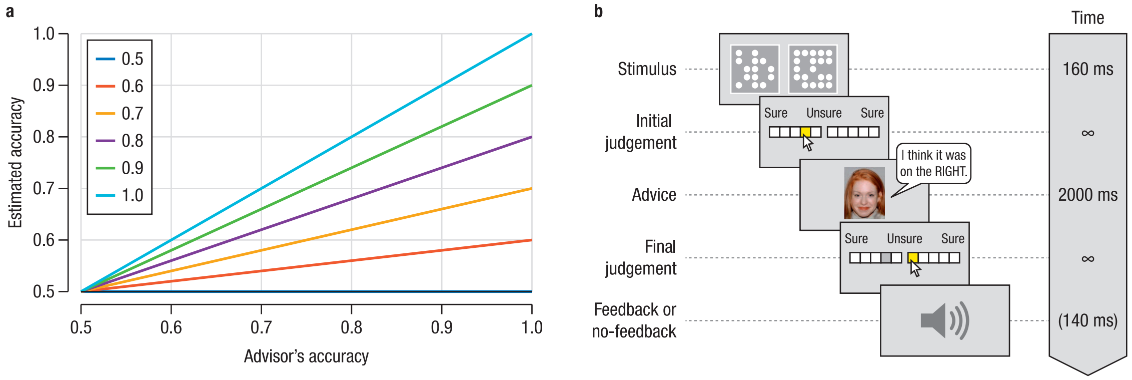

Figure 1. a: A simple model that estimates accuracy from pure agreement rate will underestimate the accuracy of any advisor unless the model itself is always correct in its judgments: Estimated accuracy = $a \times b + (1\text{-}a) \times (1\text{-}b)$, where a is the objective accuracy of the judge's decisions and b is the advisor's accuracy; that is, their likelihood of agreeing on the correct answer plus their likelihood of agreeing on an incorrect answer. b: Schematic illustration of Experiment 1 paradigm. The computerized judge–advisor system (Sniezek & Buckley, 1995) involved on each trial a perceptual decision, advice from one of four task virtual advisers, followed by a final decision. The human model in panel b is reproduced in agreement with the terms of the MacBrain Face Stimulus Set ("NimStim"). See the online article for the color version of this figure.

Larrick, & Soll, 2017; Sniezek & Buckley, 1995; Yaniv & Kleinberger, 2000). It is already well-known that confidence of both judge and advisors act as a weight in opinion aggregation: Confident judges are less likely to ask for advice and are less influenced by advice they receive (Pescetelli, Hauperich, & Yeung, 2020; Tost, Gino, & Larrick, 2012); meanwhile, confident people are trusted more and are more influential within groups and juries (Penrod & Cutler, 1995; Roediger, Roediger, Wixted, & Desoto, 2012; Swol & Sniezek, 2005; Zarnoth & Sniezek, 1997), irrespective of true accuracy (Hertz, Romand-Monnier, Kyriakopoulou, & Bahrami, 2016; Mahmoodi et al., 2015). According to a confidence heuristic (Price & Stone, 2004; Pulford, Colman, Buabang, & Krockow, 2018) an advisor's confidence signals their likely accuracy. This heuristic is normatively justified to the extent that confidence predicts objective accuracy (Bahrami et al., 2010; Henmon, 1911; Koriat, 2012), and reflects a subjective probabilistic estimate of decision accuracy (Aitchison, Bang, Bahrami, & Latham, 2015; Fleming & Daw, 2017; Meyniel, Sigman, & Mainen, 2015; Pouget, Drugowitsch, & Kepecs, 2016).

Importantly, we propose that the role of confidence goes beyond serving as an external social signal of advice accuracy (Bahrami et al., 2010; Price & Stone, 2004) and also serves as a learning signal by which an advisee can evaluate the accuracy of the advisor themselves, thus enabling formation of stable representations of advisor competence. In contrast with other strategies based on classification variability (Weiss & Shanteau, 2003), which require repeated observations, the agreement-in-confidence heuristic enables people to learn an advisor's accuracy even in one-shot interactions as a basis for weighting their advice in future interactions. This strategy is consistent with various social psychological phenomena, including the "false consensus" effect, naïve realism and social judgment theory's "latitude of acceptance," which all indicate a tendency for people to discount disagreeing opinions, underweight advice as a function of distance from one's own opinion, and consider one's own opinions as more objective or frequent than others' (Ecken & Pibernik, 2016; Liberman, Minson, Bryan, & Ross, 2012; Minson, Liberman, & Ross, 2011; Ross, Greene, & House, 1977; Schultze, Gerlach, & Rittich, 2018; Sherif, Sherif, & Nebergall, 1965; Soll & Larrick, 2009; Yaniv, 2004). Importantly, however, our approach differs in suggesting that these phenomena are part of a normatively justified strategy that enables people to discern advisors' features without the benefit of feedback: When a judge and advisor are independent, their rate of agreement varies as a simple monotonic function of their respective accuracies, so that agreement rate can be used to infer an advisor's accuracy. However, a judge using this strategy will always underestimate the accuracy of the advisor, unless they themselves are perfectly accurate (Figure 1A).[1] Using the agreement-in-confidence heuristic, a judge can have more nuanced and accurate assessments of advisors, because they can weight their learning about advisors according to their certainty in their initial view. Here we test the hypothesis that people use this agreement-in-confidence heuristic to learn about the accuracy of their advisors. Moreover, we extend these ideas to identify the boundary conditions that determine whether use of this strategy is adaptive or maladaptive depending on features of the environment, as predicted by adaptive rationality theories (Gigerenzer & Selten, 2002; Simon, 1972).

Overview of Research

In a series of experiments and agent-based modeling simulations, we explore the following three key hypotheses:

> *H1:* People can learn about the competence of their advisors even in the absence of feedback or contextual cues.

[1] Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

> *H2:* This learning depends on an agreement-in-confidence heuristic that is normatively prescribed but can lead to systematic biases in learning.

> *H3:* Use of the agreement-in-confidence heuristic by individual decision makers can influence the evolution of trust and belief within a social network.

We test the first two hypotheses—that people can learn the accuracy of advisors even in the absence of feedback, and do so based on the use of internal metacognitive information—using a judge-advisor system task (Bonaccio & Dalal, 2006). Participants (judges) performed a series of simple perceptual judgments, first giving an initial response and associated confidence, then revising decision and confidence after having been shown the opinion of different virtual advisors with predetermined informational profiles. Crucially, we manipulated across participants the presence of objective trial-by-trial feedback. The group receiving trial-by-trial feedback provides a baseline where we expected participants to correctly judge advisor accuracy (Behrens et al., 2008). The behavior of interest is whether, in the absence of feedback, participants still learn to trust advisors differentially as a function of their objective accuracy. Two measures of perceived competence were recorded to allow for dissociations between explicit and implicit behaviors: explicit numerical ratings of advisor competence versus measurement of the influence of their advice on participants' decisions (Bonaccio & Dalal, 2006). We label the influence measure as "implicit" because participants were not explicitly prompted to report it.

Experiment 1 tested whether people learn crucial characteristics of advisors, like their accuracy and calibration, in the absence of feedback (H1). When feedback is readily available, both accuracy (Behrens et al., 2008) and calibration (Tenney, MacCoun, Spellman, & Hastie, 2007) have been shown to be valued advice features affecting both trustworthiness and influence. We replicate these findings and extend them to show that people are capable of learning these advisor features even without objective feedback.

In Experiment 1, advisors' judgments were independent from the participant's judgments. Experiments 2 and 3 extended Experiment 1 to explore crucial boundary conditions of consensus and confidence-based estimates of advisor competence that are a key implication of the agreement-in-confidence heuristic (H2). When judge and advisor opinions are correlated rather than independent, agreement-based heuristics will systematically overestimate the accuracy of advisors. It is plausible that such correlations are common in real world scenarios where people share the same information, for example via news sources or social groupings, or approach questions with similar biases (Del Vicario, Bessi, et al., 2016; Sunstein, 2001; Tversky & Kahneman, 1974). To mimic these scenarios of correlated opinions, we break the coupling between agreement and accuracy by creating dependence between participants' initial judgments and the advice they receive. We show that this leads to predictable distortions in rated competence and influence, which differ from those seen when feedback is provided.

These distortions are potentially relevant to understand decision-making in many real-world environments—from online debates to information consumption—that are characterized by multiple judges and correlated information among them (Del Vicario, Bessi, et al., 2016; Kao & Couzin, 2014; Kao, Miller, Torney, Hartnett, & Couzin, 2014; Sunstein, 2001; Yaniv, Choshen-Hillel, & Milyavsky, 2009). In the final section, we use agent-based modeling to generalize the simple mechanisms of trust formation we identify empirically to consider their impact in larger groups of interacting agents. In our simulations, we show that, under realistic assumptions, the empirically identified heuristics lead to distinct patterns of trust and belief formation at the collective level (H3). Specifically, our agent-based models suggest that Bayesian normative heuristics can lead to the emergence of clusters of individuals sharing similar biases that are persistent over time.

Experiment 1

Experiment 1 investigated whether people can learn about advisors' features—specifically, their accuracy and calibration—in environments that lack objective feedback. Four virtual advisors with differing profiles were designed, who differed in accuracy and calibration (Fleming & Lau, 2014). Participants repeatedly experienced each advisor to give them the opportunity to learn about the quality of their advice. The crucial question of interest was whether the perceived competence of advisors would be sensitive to their objective accuracy and calibration, even if participants did not have access to feedback on each trial. By including a second group of participants, who did have access to trial-by-trial feedback, we could also assess the effect of using objective versus subjective learning signals on evaluations of advisor competence.

Method

Participants. There were 46 participants (26 females, age = 23 ± 0.45), half of whom were pseudorandomly assigned to the feedback condition and the other half to the no-feedback condition.

Paradigm. The perceptual task (Boldt & Yeung, 2015) required participants to judge which of two briefly presented boxes contained more dots (Figure 1B). One box contains more dots, specifically $ndots = 200 + d$, compared to the other with $ndots = 200 - d$, with dots pseudorandomly assigned to locations in a 20×20 grid anew on each trial and with equal trial numbers with left and right box as the correct answer. By manipulating the d parameter, we titrated the difficulty of the task (Treutwein, 1995) to ensure similar overall accuracy across participants (nominal accuracy rate = 70.7%).

Participants registered their response and confidence judgment, unspeeded, by mouse-click on a semicontinuous scale in 10 steps, ranging from "100% sure left" to "100% sure right." Text landmarks signaling 10% increases aided the interpretation of the scale. The middle point of the scale (50% or total uncertainty) was removed and a gap appeared instead, meaning that participants had to commit to one interval (two-alternative forced-choice). After confirming their response with the spacebar, one of four different advisors appeared centrally as a head-shot picture. Advice was provided in the form of spoken sentences that expressed a binary level of confidence (low vs. high) and either agreement or disagreement with the participant's judgment (see the online supplementary material for further details).

Participants were then given the opportunity to update their decision and confidence level using the same interface and input

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

method as used in the preadvice period. In the feedback condition only, after the final decision was confirmed, a high frequency tone indicated whenever the participant's final decision was incorrect. In the no-feedback condition, a new trial started immediately after participants had confirmed their final answer.

At the end of each block, all participants saw a summary of their postadvice percentage accuracy. Because advisors appeared equally often and in randomized order within blocks, this feedback could not favor one advisor over the others. Participants performed 500 trials across 10 experimental blocks. Prior to these, two initial blocks with a fifth advisor served as practice and were removed from all the analyses. On each experimental block, each advisor appeared 10 times. Ten randomly selected trials within each block were presented with a black silent silhouette and a postadvice decision was not required (null trials), to motivate participants to provide meaningful answers in their preadvice answers on each trial. After every two experimental blocks, participants answered a brief questionnaire about their explicit opinions about the four advisers. Four questions asked participants to directly rate on a scale from 1 (*not at all*) to 50 (*extremely*) how much they thought each adviser was accurate (Q1), confident (Q2), trustworthy (Q3), and influential on their own choices (Q4; see the online supplementary material for complete description).

Manipulation. We orthogonally manipulated the average accuracy of the four advisors and their confidence-to-accuracy calibration. The two accurate advisors gave correct answers on 80% of trials, whereas the two inaccurate advisors gave correct answers 60% of trials. Crossed with this factor, the two calibrated advisors were always correct when expressing answers with low uncertainty ("I'm sure") and less accurate when uncertain ("I think"), whereas the two uncalibrated advisors expressed uncertainty independently from objective accuracy. This led to the profiles shown in Table 1. Note that all advisors were equally often confident versus unconfident across trials, to avoid participants simply trusting the advisor who was the most confident *on average* when objective feedback is not available (Sah, Moore, & Maccoun, 2013).

Exclusion criteria. An exclusion criterion was set a priori for staircase convergence. Participants who showed progressively increasing thresholds (i.e., increasing dot difference *d* across the experiment) were to be eliminated as this indicated that they were randomly guessing. None of the participants had to be removed when this criterion was applied to our sample. At the end of the experiment the average difficulty parameter *d* across participants (pooled data) was 9.6 ± 2.81.

Table 1
Experiment 1 Advisors' Profiles

	Advisors			
Events count	Accurate calibrated	Accurate uncalibrated	Inaccurate calibrated	Inaccurate uncalibrated
Incorrect confident	0	1	0	2
Incorrect unconfident	2	1	4	2
Correct unconfident	3	4	1	3
Correct confident	5	4	5	3

Note. Values in the central section represent the number of times each event occurred over the course of a 50-trial block (10 null trials, with no advisor, are not shown). Calibration (metacognitive sensitivity) of each advisor and their informativeness are reported in the online supplementary material.

Results

The key set of analyses assessed whether participants were sensitive to advisors' features (accuracy and calibration) in the absence of feedback (i.e., the between-participants manipulation). These questions were investigated through the analysis of both explicit ratings of perceived competence and implicit influence measure, with data from all blocks collapsed given that preliminary analyses indicated no notable effects of time. A summary table is provided in the Appendix, summarizing all analysis of variance (ANOVA) results of rated competence (Table A1) and influence (Table A2), for all three experiments.

Competence ratings. Participants provided explicit ratings of advisor competence after every second experimental block, along four dimensions: accuracy, confidence, trustworthiness, and influence. An initial rating was provided before the start to assess baseline perceived competence, for example, due to advisor appearance. Baseline ratings were then subtracted from subsequent ones to remove these effects. Ratings were then converted into a unitary measure of perceived competence via principal components analysis (see the online supplementary material). We used PCA over simple averaging because (a) we were agnostic on which dimensions of advisor competence were made salient by our manipulation and (b) we did not want to average over distinct constructs (e.g., accuracy and trust).

A mixed-design ANOVA on the resulting rated competence scores, with factors of feedback (between-participants s) and advisor accuracy and calibration (both within-participant s), revealed significant main effects for advisor accuracy, $F(1, 44) = 9.68$, $p = .003$, $\eta^2_G = 0.079$, and advisor calibration, $F(1, 44) = 12.32$, $p = .001$, $\eta^2_G = 0.076$, but not feedback ($F < 1$). Participants gave higher competence ratings for accurate over inaccurate advisors, and for calibrated over uncalibrated advisors. No interaction term reached significance, $F(1,44) < 1.9$, $p > .16$. Importantly, neither within-participants manipulation interacted with Feedback, suggesting that participants were sensitive to the accuracy of the advice and that this sensitivity did not vary consistently according to the presence or absence of feedback (Figure 2A).

We next ran planned two-way ANOVAs separately for each feedback group to assess whether the overall patterns described above were also reliable in each group. In the feedback group this analysis revealed reliable main effects of both advisor accuracy, $F(1, 22) = 8.26$, $p = .008$, $\eta^2_G = .09$, and advisor calibration, $F(1, 22) = 8.71$, $p = .007$, $\eta^2_G = 0.12$, but no reliable interaction, $F(1, 22) = 1.24$, $p = .27$, $\eta^2_G = .02$. In the no feedback group, although comparable numerical trends were apparent, neither main effect of accuracy, $F(1, 22) = 3.42$, $p = .07$, $\eta^2_G = 0.06$, and calibration, $F(1, 22) = 4.03$, $p = .05$, $\eta^2_G = 0.04$, reached statistical significance. The interaction term was not significant ($F < 1$).

Influence. Influence was quantified as the signed difference between post and preadvice confidence (Equation S3 in the online supplementary material) and represents the shift in a participant's expressed judgment observed after social information. We replicated the results separately for agreement and disagreement trials (Section 2 in the online supplementary material). Influence results

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

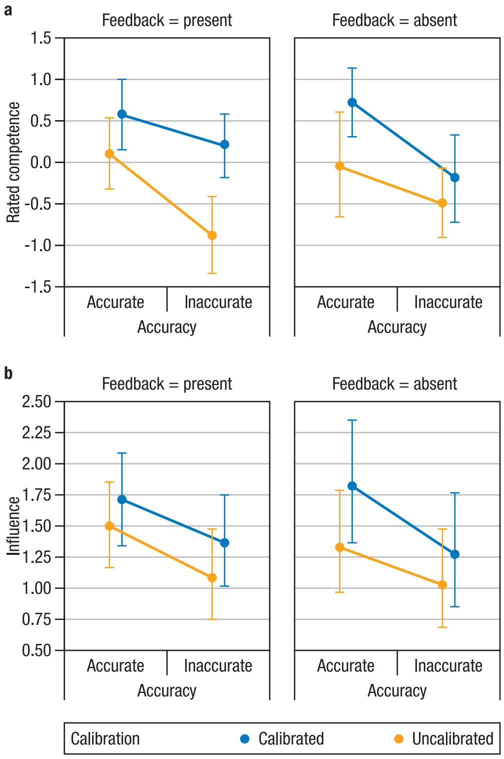

Figure 2. Experiment 1: Rated competence and influence patterns for humans. a: Average competence ratings in the two feedback groups as a function of advisor accuracy and calibration. Error bars represent 95% bootstrap confidence intervals. b: The effect of advisor accuracy, advice confidence and calibration, and feedback group on the influence measure, averaging across the two levels of advisor confidence. Error bars represent 95% bootstrap confidence intervals. See the online article for the color version of this figure.

(Figure 2B) were analyzed using a mixed-design ANOVA that included the same factors of feedback presence (between-participants), advisor accuracy and calibration (both within-participant) as above, together with the additional within-participants factor of the confidence expressed by the advisor on a given trial. This analysis revealed significant main effects of advisor accuracy, $F(1, 44) = 14.80$, $p < .001$, $\eta_G^2 = 0.02$, and calibration, $F(1, 44) = 15.84$, $p < .001$, $\eta_G^2 = 0.01$, mirroring the pattern of results seen for rated competence. A reliable main effect of advisor confidence, $F(1, 44) = 55.82$, $p < .001$, $\eta_G^2 = 0.12$, unsurprisingly indicated that more confidently expressed advice had greater influence. There was a significant interaction between calibration and advisor confidence, $F(1, 44) = 9.62$, $p = .003$, $\eta_G^2 = 0.004$, indicating that the effect of confidently (vs. unconfidently) expressed advice was greater for calibrated advisors than uncalibrated advisors. The analysis also revealed a significant three-way interaction between advisor accuracy, calibration, and confidence, $F(1, 44) = 4.75$, $p = .03$, $\eta_G^2 = 8.5e - 04$, indicating that the two-way interaction between calibration and confidence was larger for inaccurate advisors than accurate ones. importantly, there was no reliable main effect of feedback ($F < 1$), nor any reliable interaction between feedback and any other main effects, all Fs(1,44) < 1.9, $ps > .29$, suggesting again that participants' sensitivity to advisor accuracy, here expressed in terms of the influence of their advice, did not depend significantly on the provision of trial-by-trial feedback.

Although we observed no reliable effects of feedback, we ran planned follow-up ANOVAs for each feedback group separately to assess whether the overall patterns described above were also reliable in each group. These analyses revealed significant effects in both groups of advisor accuracy, Fs(1,22) > 1.9, $ps < .05$, $\eta_G^2 > 0.02$, and calibration, Fs(1,22) > 6.35, $ps < .01$, $\eta_G^2 = .01$, and advice confidence, Fs(1,22) $= 22.95$, $ps < .001$, $\eta_G^2 = 0.07$. For the feedback group, we also found a reliable two-way interaction between calibration and confidence, Fs(1,22) $= 7.37$, $p = .01$, $\eta_G^2 = .008$, and a reliable three-way interaction between accuracy, calibration, and confidence, $F(1,22) = 8.32$, $p = .008$, $\eta_G^2 = .003$, indicating that the influence difference between calibrated and uncalibrated advisors was mainly shown for highly confident advice compared to uncertain advice, and more so for inaccurate advisors than accurate ones. Similar patterns were seen numerically in the no feedback group, but the effects were not statistically reliable, Fs(1,22) < 2.59, $ps > .12$.

Discussion

Overall, participants in Experiment 1 gave higher ratings of competence for, and were more influenced by, advisors who were characterized by high accuracy rates and high calibration. The finding that advisor confidence and calibration affect perceptions of advice is broadly consistent with previous findings (Price & Stone, 2004; Sah et al., 2013; Sniezek & Van Swol, 2001; Tenney et al., 2007). However, the present results extend this work to show, in a highly controlled perceptual task, that participants distinguish advice along these dimensions even in the absence of objective feedback. Thus, although rated competence and influence were not identical across groups, no consistent (statistically reliable) differences were observed, and overall participants in the no feedback group were able to learn distinguishing characteristics of advisors (accuracy and calibration).

These empirical results are consistent with the hypothesis that, in the absence of feedback, people make use of internal signals to evaluate the quality of advice they receive. In algorithmic simulations based on straightforward implementations of two related learning strategies—estimating advice accuracy according to simple agreement, or agreement-in-confidence—we show how this differentiation of advice quality can be achieved (Section 3 in the online supplementary material). As such, the combined empirical and simulation results indicate the normative value of these heuristic strategies in enabling people to discern the usefulness of advice in feedback-poor environments: When advice is independent from a judge's initial opinion, agreement and confidence covary with accuracy and thus are useful cues to integrate over time so to learn about the competence of an advisor. Extending these ideas, Experiments 2 and 3 aimed to test the agreement-in-confidence hypothesis more directly, and to explore crucial limitations in the use of agreement and confidence in estimating advisor accuracy, by exposing participants to advice that was not independent of their own initial decisions.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Experiment 2

Experiment 1 showed that people are able to detect subtle advisor differences even in the absence of feedback. According to our hypothesis, they achieve this differentiation of advisor competence using an agreement-in-confidence heuristic, whereby feedback is replaced by the interaction of past agreement with the advisor and internal metacognitive signals: Specifically, to the extent that an observer's internal confidence is calibrated (i.e., predictive of their objective accuracy), they can validly learn about advisor's accuracy (vs. inaccuracy) by integrating agreement (vs. disagreement) with their own confidently held opinions.

Giving weight to advice that agrees with one's own view seems normatively appropriate from a Bayesian standpoint (Dawes, 1989; Krueger & Clement, 1994), but depends critically on the assumption that observers are independent. However, people's judgments are rarely independent and distorted only by random noise (Koriat, 2012; Krause, Ruxton, & Krause, 2010): Dependence between individuals' opinions can arise from use of similar cognitive heuristics that lead to similar reasoning errors or information sampling (Tversky & Kahneman, 1974; Vandormael, Herce Castañón, Balaguer, Li, & Summerfield, 2017), from being exposed to similar signals (Kao & Couzin, 2014; Kao et al., 2014) or belonging to the same social clique (Jamieson & Cappella, 2008; Jasny, Waggle, & Fisher, 2015; Sunstein, 2001). People tend not to take into account advisors' judgments interdependence (Yaniv et al., 2009). As a result, crowds are known to be susceptible to error cascades (Le Bon, 1895; Mackay, 1841), economic bubbles (De Martino, O'Doherty, Ray, Bossaerts, & Camerer, 2013), polarization (Myers & Lamm, 1976), and groupthink (Janis, 1972; Turner & Pratkanis, 1998). Moreover, in the context of advice, advisors' opinions and suggestions might be deliberately framed in relation to the advisee. For any or all of these reasons, advice may not be independent of a decision maker's judgment in many realistic scenarios. Heavily weighing agreeing advisors might in these cases be maladaptive. As a first exploration of scenarios with correlated advice, we thus investigated whether people rely on accurate advisors or instead advisors who tend to agree with their own initial judgment.

In this experiment, advisor accuracy was manipulated so that two advisors were, on average, highly accurate (around 80% accuracy) and two advisors were, on average, relatively inaccurate (around 60% accuracy). Orthogonally, advisor agreement rate with the participant's initial judgment was manipulated to create two advisors who agreed with the participant frequently (around 80% of trials) and two advisors who tended to have a lower agreement rate with the participant (around 60%). We conceive of an advisor with low accuracy but a high rate of agreement as someone who shares biases with the participant and so makes similar (correlated) mistakes. The accurate but disagreeing advisor conversely represents an advisor who uses different information and therefore tends to be correct when the participant makes mistakes, and vice versa. Of interest was the impact on the perceived competence and influence of each advisor, as a function of the advisors' accuracy and agreement rates and, separately, the participants' access to objective feedback.

Method

Participants. The experiment included 46 participants, equally divided between the two feedback groups (37 females in total, 18 of whom were in the Feedback group, $M_{\text{age}} = 21.63 \pm 3.02$).

Paradigm. The overall design was very similar to Experiment 1, with advice provided by four virtual advisors, characterized by distinct informational profiles, appearing in the context of the same dot-count perceptual decision task. Advice was always presented in the form of a binary left/right judgment (i.e., with no accompanying indication of high vs. low advice confidence), which could agree or disagree with participants' original judgments. Participants completed 10 blocks of 44 trials each. The presentation of the four advisors was randomly shuffled across trials within-block, with each advisor appearing exactly 10 times. Four additional trials served as null trials (as above). Explicit ratings of advisor competence along four dimensions were again collected at the end of every second block and aggregated as before: accuracy (Q1), likability (Q2), trustworthiness (Q3), and influence (Q4).

Manipulation. To disentangle advisor agreement rate and accuracy, the probability of agreement conditional on the participant's choice accuracy was manipulated. Through the staircase procedure, it was expected that all participants would converge to an accuracy level of about 70%. This enabled us to manipulate advisor accuracy and agreement rate separately, by predetermining the probability of agreement differently according to whether the participant's initial decision was correct versus incorrect. Both accuracy and agreement were manipulated to have two levels (high = 80% and low = 60%). This gave rise to the four advisor profiles defined in Table 2. Probabilities are expressed as a fraction over the number of participants' expected correct (seven) and incorrect (three) judgments, over the number of encounters with

Table 2
Experiment 2 Advisors' Profiles

Advisor profile	High accuracy, high agreement	High accuracy, low agreement	Low accuracy, high agreement	Low accuracy, low agreement
$p(Agr \mid Correct_s)$	6.5/7	5.5/7	5.5/7	4.5/7
$p(Agr \mid Incorrect_s)$	1.5/3	0.5/3	2.5/3	1.5/3
Expected accuracy rate	80%	80%	60%	60%
Expected agreement rate	80%	60%	80%	60%

Note. Expected accuracy and agreement rates of different advisors are disentangled by manipulating the probability of the advice agreeing with the participant, conditional on the participant's accuracy. Probabilities are expressed as a fraction of the number of participants' expected correct (7) and incorrect (3) judgments, during the number of encounters with one advisor (10) in a single experimental block. We report each advisor information value in online supplemental material.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

one advisor during one block (10). Of key interest was the separate impact of advisor accuracy and agreement rate on explicit ratings of competence and implicit measures of advisor influence, separately for the feedback and no feedback groups.

Exclusion criteria. The first two experimental blocks were removed from the analysis to allow the staircase procedure to fully adapt to each individual's threshold. This was necessary given that our manipulation was heavily dependent on the expected accuracy rate of the participants. A further exclusion criterion was set to exclude all participants whose threshold never converged, which suggests a random response strategy. None of the participants had to be removed on the basis of this criterion. The perceptual task difficulty *d* (dot difference between boxes) after staircasing was 9.93 ± 2.96 (pooled data).

Results

As for Experiment 1, separate analyses were conducted on rated competence and influence measures. Of interest was whether the manipulated within-participants factors (advisor agreement rate and accuracy) varied in impact across two groups of participants, with differential access to objective trial-by-trial feedback.

Competence ratings. A mixed-design ANOVA was run on competence ratings with feedback group as a between-participants factor and advisor accuracy (low vs. high) and agreement rate (low vs. high) as within-participants factors. This analysis revealed significant main effects for both accuracy, $F(1,44) = 8.36$, $p < .005$, $\eta^2_G = .06$, and agreement rate, $F(1,44) = 22.52$, $p < .001$, $\eta^2_G = .1$, but not for feedback ($F < 1$). A significant interaction between feedback and accuracy, $F(1,44) < 8.41$, $p = .005$, $\eta^2_G = .06$, indicated much greater impact of advisor accuracy on participants' rated competence when feedback was available than when it was absent (Figure 3A). There was no reliable interaction between feedback and agreement rate, $F(1,44) = 1.88$, $p = .17$, $\eta^2_G = .01$, or between agreement and accuracy ($F < 1$), nor a significant three-way interaction ($F < 1$).

Separate planned ANOVAs were conducted for the two groups separately. Analysis of competence ratings from the feedback group revealed significant effects of both advisor's accuracy, $F(1,22) = 21.36$, $p < .001$, $\eta^2_G = .20$, and agreement rate, $F(1,22) = 5.62$, $p = .02$, $\eta^2_G = .06$, but no significant interaction ($F < 1$). The effect of accuracy was much stronger than the one observed for agreement as indicated by the generalized eta squared values (Bakeman, 2005). Nonetheless, agreement rate had a significant effect on rated competence even when controlling for objective accuracy. Analysis of competence ratings from the no-feedback group found that agreement rate, $F(1,22) = 18.91$, $p < .001$, $\eta^2_G = .18$, but not accuracy ($F < 1$) affected participants' explicit competence ratings, with no significant interaction ($F < 1$). These findings suggest that when feedback was not directly available to estimate partners' accuracy, participants rated agreeing advisors as more competent than disagreeing ones but did not consistently differentiate accurate versus inaccurate advisors.

Influence. The next analysis focused on influence, our implicit measure of perceived competence, using the same mixed-design ANOVA as above. Table A2 in the Appendix summarizes the results of this ANOVA and corresponding analyses in Experiments 2 and 3. This analysis revealed significant main effects of advisor accuracy, $F(1,44) = 14.79$, $p < .001$, $\eta^2_G = .04$, and agreement rate, $F(1,44) = 13.91$, $p < .001$, $\eta^2_G = .03$, with no significant interactions, including between accuracy and feedback ($F < 1$) and between agreement rate and feedback, $F(1,44) = 2.01$, $p = .16$, $\eta^2_G = .005$. Thus, participants were more influenced by accurate advisors and by advisors characterized by high agreement rates with their own judgments, and these effects did not vary consistently as a function of whether or not trial-by-trial feedback was available. We replicate the results for agreement and disagreement confidence changes separately (Section 2 in the online supplementary material).

Planned ANOVAs on the influence measure for each group separately revealed, for the feedback group, significant effects of advisor accuracy, $F(1,22) = 12.71$, $p = .001$, $\eta^2_G = .06$, and agreement rate, $F(1,22) = 5.81$, $p < .02$, $\eta^2_G = .01$, with no significant interaction ($F < 1$). As would be expected, when feedback was available, participants were more influenced by accurate advisors over inaccurate ones. More surprisingly, they were also more influenced by advisors who more often agreed with their own judgment, even though feedback was available that indicated equivalent accuracy rates for pairs of advisors characterized by different agreement rates. Analysis of advisor influence in the no-feedback group revealed a similar pattern, with significant main effects for both advisor agreement rate, $F(1,22) = 8.60$,

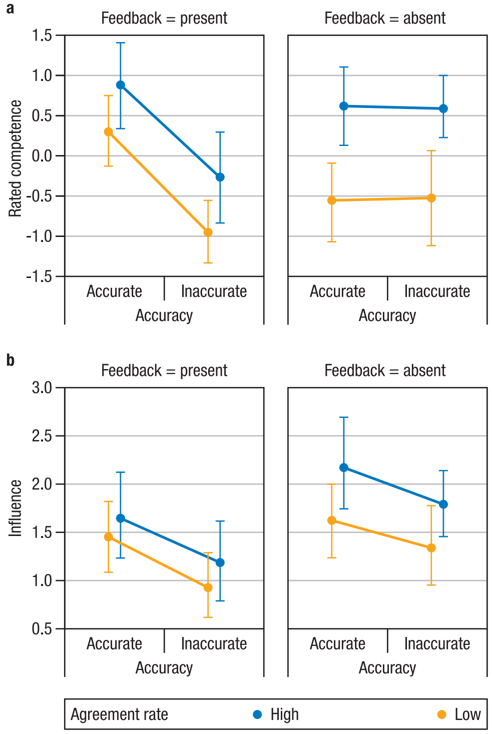

Figure 3. Experiment 2: Rated competence ratings and influence for human participants and simulations. a: Average competence ratings in the two feedback groups, divided by advisor accuracy rate and agreement rate. Error bars represent 95% bootstrap confidence intervals. b: Influence that advice had on participants' opinions in the two feedback groups and divided by advisor accuracy rate and agreement rate. Error bars represent 95% bootstrap confidence intervals. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

$p = .007$, $\eta^2_G = .06$, and accuracy, $F(1,22) = 4.09$, $p < .05$, $\eta^2_G = .02$, although the effect size of agreement was the greater of the two. No significant interaction was observed ($F < 1$). Thus, in the no feedback group, the results suggest a dissociation between what people reported in explicit competence ratings, with higher competence reported for agreeing advisors regardless of accuracy, and what was apparent in advisors' influence on participants' decisions, which showed effects of both accuracy and agreement rate.

Discussion

The results of this experiment reveal a strong effect of advisor agreement on perceived competence, regardless of the presence or absence of objective feedback and of the objective accuracy of advice: Participants showed greater reliance on advisors who more often agreed with their own initial judgments, both in their explicit ratings of competence and in the degree to which the advisors influenced their final decisions. This impact of agreement is in line with established findings in social psychology—for example, phenomena of naïve realism and latitude of acceptance—that people tend to discard disagreeing opinions and tend to see their own subjective views as being more objective than others' (Liberman et al., 2012; Minson et al., 2011; Sherif et al., 1965; Shultz, Katz, & Lepper, 2001). Our findings indicate that these effects extend to perceived competence that is learnt through repeated interaction with different advisors, and moreover that it persists even when participants have access to objective feedback. The agreement effect apparent in the feedback group—not predicted by a simple accuracy heuristic that learns advisor accuracy based on observed feedback (Section 3 in the online supplementary material)—suggests an intrinsic value of agreement when learning about advisor competence.

When feedback was unavailable, participants' reliance on advisors was dominantly determined by advisors' agreement rate, with much weaker effects of the objective accuracy of their advice. Nevertheless, at least for the implicit measure of advisor influence, we found that participants were still able to distinguish more accurate from less accurate advisors. Importantly, going beyond previous findings, this result cannot be explained in terms of participants simply downweighting advice on a given trial when the advice disagrees with their initial opinion (Liberman et al., 2012; Sherif et al., 1965). Rather, we observe patterns of trust and influence that reflect learning across aggregated sets of trials, such that advisors who more regularly disagreed with a participants' choices were less influential, even on those occasions where their advice happened to agree with the participants' view. Conversely, we find that advisors who more regularly agreed with the participant were more influential, even on those occasions where their opinion diverged from the participants' initial choice.

The subtle but consistent effect of advisor accuracy—even when feedback was absent and agreement rates were matched—is predicted by our agreement-in-confidence heuristic, which learns about the trust of an advisor by accumulating agreement instances weighed by internal decision confidence (Section 3.6 and Figure S6 in the online supplementary material). The absence of this effect in explicit competence ratings is the only instance we found in our experiments of a systematic divergence between explicit ratings of competence and implicit measures of influence, which generally produced very similar patterns of results. The dominance of agreement over accuracy in explicit ratings might have resulted in a halo-dumping effect (Clark & Lawless, 1994) whereby, when prompted to discriminate among advisors, participants used only the most accessible dimension (i.e., agreement rate). In Experiment 3, we therefore matched all advisors in terms of agreement and accuracy rates, while at the same time varying the amount of shared information and bias between the advisors and the participants, to provide a direct test of the hypothesis that internal confidence judgments are used when making inferences about the competence of others.

Experiment 3

Experiment 3 aimed to provide stronger evidence that subjective confidence contributes to the formation of judgments about advice accuracy. Here, we manipulated the probability that advice would agree with a participant's initial judgment, conditional on their initial confidence in that judgment. There were three advisors: (a) an unbiased advisor who tended to agree with the participant's choice about 70% of the time, independent of participants' confidence; (b) a "bias-sharing" advisor who more often agreed with the participant's initial choice when the participant expressed high confidence in this choice; and (c) an "antibias" advisor who more often agreed with the participant when the participant was unsure in their initial decision. Crucially, overall agreement rate and accuracy was identical across the three advisors. The labeling of the advisors in terms of bias does not reflect an actual bias manipulation, but rather reflects our aim to capture a property that may hold in many real-world situations, where the individuals share biases in their opinions and choices (e.g., reflecting their shared reliance on common sources of information).

Method

Participants. Fifty participants were tested and divided in the two experimental groups. Because of participants failing to attend or complete sessions, numbers across groups were unbalanced with 24 participants in the no-feedback group and 26 in the feedback group.

Paradigm. The experiment consisted of 12 experimental blocks of 35 trials each, with each of the three advisors seen on 10 trials and with 5 null trials. The perceptual task was the same as for Experiments 1 and 2, except that a different confidence rating scale was used because the 10-point scale used in Experiments 1 and 2 would not allow us to distinguish fine gradations in confidence needed here. Instead, participants rated their confidence on a 100-point scale (50 points per interval, left vs. right). Two blocks of 25 trials served as the practice blocks and used a fourth practice advisor. Questionnaires were again administered every two blocks.

Manipulation. The advice profiles of the three advisors were manipulated so that advisors were matched for accuracy (70%) and their overall agreement rate (70%) with participants' initial decisions. The pattern of agreement was manipulated, however, such that the three different advisors' likelihood of agreement varied according to the participant's initial confidence (Table 3 and Figure 4). To this end, the distribution of the participant's preadvice confidence judgments was divided into three confidence bins: the low, middle, and high confidence bins, comprising 30%, 40%,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Experiment 3 Advisors' Profiles

	Advisors		
Advisor profile	Bias-sharing	Unbiased	Anti-bias
$p(Agree \mid Correct^s, Confidence^s_{low})$	60%	70%	80%
$p(Agree \mid Correct^s, Confidence^s_{mid})$	70%	70%	70%
$p(Agree \mid Correct^s, Confidence^s_{high})$	80%	70%	60%
$p(Agree \mid Incorrect^s)$	30%	30%	30%

Note. Agreement probability of different advisors is manipulated conditional on the participant's preadvice confidence and accuracy. This manipulation allowed to create three different advisors who were matched in terms of agreement rate and accuracy, but who differed in terms of information value (see Section 1.2 in the online supplementary material for details).

and 30% of trials, respectively. On trials in which the participant's initial judgment was correct, the three advisors had different agreement patterns across these bins. An unbiased advisor had a probability of agreement of 70% independent of the participant's confidence. A bias-sharing advisor had an 80% probability of agreeing when the participant was highly confident and 60% when the participant expressed low confidence in their initial decision. An antibias advisor, conversely, had 60% probability of agreement when the participant was highly confident and 80% when s/he was uncertain. All three advisors had equal chance of agreement when the participant's decision was correct and preadvice confidence fell in the middle bin (70%). Likewise, all advisors had a 30% agreement rate independent of the participant's confidence when the participant's initial decision was incorrect. This ensured that all advisors were matched across all trials in terms of average agreement rate ($0.7 * 0.7 + 0.3 * 0.3 = 0.58$) and accuracy ($0.7 * 0.7 + 0.3 * 0.7 = 0.7$).

By limiting analyses to trials within the intermediate confidence bin, we could compare advisors on trials that were matched for confidence and a priori likelihood of advice agreement. The confidence reference distribution used to assign trials to the low, middle, and high confidence bins was first set up on the basis of each participant's confidence ratings in the first two practice blocks. The reference distribution was updated after each block to reflect the distribution of confidence judgments provided during the previous two blocks, to allow for possible shifts of confidence during the course of the experiment.

Following Experiment 2, we expected different patterns of results to emerge from feedback and feedback-free conditions. If people use subjective confidence to learn about advisors' accuracy (confidence model in Section 3 in the online supplementary material), we should observe that competence ratings and influence will favor the bias-sharing advisor over the other advisors when feedback is unavailable. This is because the participant will experience more high-confidence agreements and fewer low-confidence agreements with the bias-sharing advisor compared to the other two. In contrast, heuristics using simple agreement counts or objective feedback (consensus and accuracy models in Section 3 in the online supplementary material) would not distinguish among advisors, given that advisors are matched in terms of objective accuracy and agreement rates.

Exclusion criteria. An exclusion criterion based on staircase convergence was set so to exclude all participants who showed random guessing. Application of this criterion resulted in the exclusion of one participant from the feedback group and one participant from the no-feedback group, leaving a total of 25 and 23 participants in these groups, respectively. Average difficulty parameter d was 9.98 ± 2.82 (pooled data).

Results

Degrees of freedom were corrected for violations of sphericity according to the Greenhouse-Geisser procedure, with epsilon values reported as appropriate.

Competence ratings. A mixed-design ANOVA on competence ratings from the end-of-block questionnaires revealed no significant main effect of advisor type, $F(2,92) = 1.66$, $p = .19$, $\eta^2_G = .03$, $\epsilon = 0.99$, or feedback ($F < 1$), but a significant interaction between these factors, $F(1,92) = 6.64$, $p < .002$, $\eta^2_G = .12$, $\epsilon = 0.99$. Figure 5A shows how advisor perceived competence varied according to the presence versus absence of feedback.

Planned follow-up one-way ANOVAs for each group separately revealed, for the feedback group, a significant effect of advisor type, $F(2,48) = 4.90$, $p = .01$, $\eta^2_G = .16$, with rated competence being highest for the antibias advisor, intermediate for the unbiased advisor, and lowest for the bias-sharing advisor. Pairwise comparisons indicated that the bias-sharing advisor was perceived significantly less accurate than the antibias advisor, $t(24) = 3.09$, $p = .004$, $d = 1.07$, with no reliable differences observed otherwise, $ts(24) < 1.60$, $p > .12$, $d < 0.56$. This pattern, with a high level of reliance on the antibias advisor and lower in the bias-sharing advisor is in accordance with the pattern of information gain of each advisor (Supplementary Table S1 in the online supplementary material), rather than advisors' objective accuracy. Similar results are reported below for influence, suggesting this pattern is robust to different measures of trust.

A corresponding analysis on competence ratings from the no-feedback group also revealed a significant difference among advisors, $F(2,44) = 3.56$, $p = .03$, $\eta^2_G = .13$, but the pattern observed

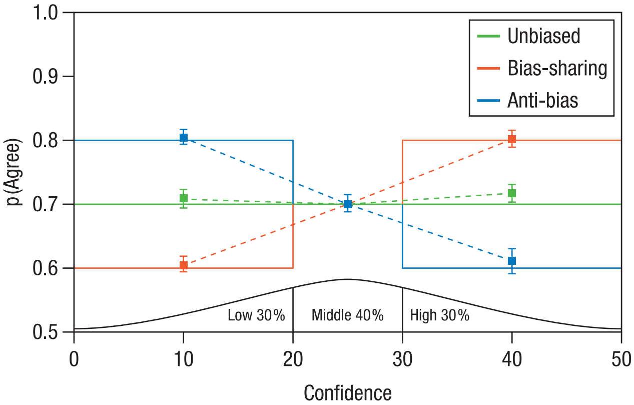

Figure 4. Experimental manipulation of Experiment 3. The probability of advisor's agreement conditional on participant's accuracy and confidence was manipulated so that the three advisors differed in their pattern of agreement (i.e., bias) despite being equal on average agreement rate and accuracy rate. Continuous lines represent expected agreement rates, dashed lines represent empirical data pooled across the two feedback groups. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

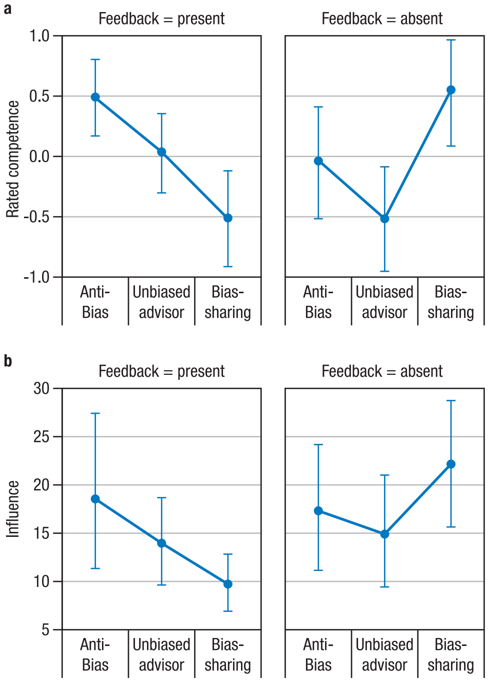

Figure 5. Experiment 3: Competence ratings and influence of human participants. a: The effect of advisor type on competence ratings, separately for the feedback and no-feedback condition. Error bars represent 95% bootstrap confidence intervals. b: The effect of advisor type on the influence measure, divided by the two feedback conditions. Error bars represent 95% bootstrap confidence intervals. See the online article for the color version of this figure.

was very different. As predicted, rated competence was highest for the bias-sharing advisor when feedback was absent, numerically so compared with the antibias advisor, $t(22) = 1.48$, $p = .07$, $d = 0.53$, one-tailed, and significantly so compared with the unbiased advisor, $t(22) = 2.81$, $p = .005$, $d = 0.98$, one-tailed. The antibias and the unbiased advisors did not significantly differ from each other ($p > .1$, $d = 0.41$), but rated competence was, unexpectedly, numerically higher for the antibias observer than the unbiased advisor.

Influence. Similar patterns of trust were apparent as measured implicitly via advisor influence (confidence change following advice). A mixed-design ANOVA on measured influence revealed no significant main effect of feedback, $F(1,46) = 1.36$, $p = .24$, $\eta^2_G = .01$, nor advisor, $F(2,92) = 1.12$, $p = .33$, $\eta^2_G = .009$, $\epsilon = 0.77$, but a significant interaction between the two, $F(2,92) = 4.80$, $p = .01$, $\eta^2_G = .03$, $\epsilon = 0.77$ (Figure 5B).

When looking at influence in the feedback condition only, a one-way ANOVA revealed a marginally significant effect of advisor type, $F(2,48) = 2.88$, $p = .06$, $\eta^2_G = .05$. Planned comparisons (two-tailed *t* tests) showed that the bias-sharing advisor was less influential than both the antibias advisor, $t(24) = 1.98$, $p = .05$, $d = 0.54$, and the unbiased advisor, $t(24) = 2.26$, $p = .03$, $d = 0.44$. The antibias advisor was numerically more influential than the bias-sharing advisor, but this difference was not reliable, $t(24) = 1.10$, $p = .28$, $d = 0.26$. Similar numerical trends were observed in both agreement and disagreement trials separately (Section 2 in the online supplementary material). The effects remained significant when considering disagreement trials only.

In the no-feedback condition, there was a significant effect of advisor, $F(2,44) = 3.25$, $p = .04$, $\eta^2_G = .03$. Planned comparisons showed that the bias-sharing advisor was significantly more influential than the unbiased advisor, $t(22) = 2.63$, $p = .007$, $d = 0.46$, one-tailed, and numerically more influential than the antibias advisor, $t(22) = 1.46$, $p = .07$, $d = 0.29$, one-tailed. No significant difference was found between the unbiased and the antibias advisors ($p > .1$, $d = 0.16$), but the direction of the difference was the same as for rated competence, with the antibias advisor somewhat more influential on participants' decisions and confidence than the unbiased advisor. Thus, these results seem to suggest that the bias-sharing advisor was more influential than the other two when trial-by-trial feedback was not available.

In summary, as in Experiment 2, we find that participants' perception of advisor accuracy varies systematically according to whether trial-by-trial feedback is present or absent when advice is correlated with their own initial decisions. In particular, advisors' perceived competence follows their relative informativeness when feedback is provided, but when feedback is absent participants tend to rely on advisors who share their judgment biases (i.e., who agree with their confidently held judgments). Of three heuristic models fitted to these empirical data—using accuracy, agreement, and agreement-in-confidence learning signals respectively—only the agreement-in-confidence heuristic was able to produce diversified beliefs about the advisors (Section 3 in the online supplementary material).

Discussion

Experiment 3 showed again that systematic differences in perceived competence emerge (both in humans and models) as a function of feedback when advisors' judgments are nonindependent from the advisee's, and further demonstrated a strong influence of our agreement-in-confidence manipulation on trust. Here, the presence of feedback partly reversed the pattern of competence ratings and influence measure that was observed when trial-by-trial feedback was unavailable: When objective feedback was provided, people perceived as more competent, and were more influenced by, the advisor who more frequently agreed with them when they themselves were unsure (vs. less frequently when they were sure), and showed less trust in an advisor who tended to agree with them in decisions in which they were already sure. Evidently, though advisor accuracy and agreement are critical determinants of perceived competence (as evidenced in Experiment 2), participants remain sensitive to other dimensions of advice. Here, they appeared sensitive to the informational value (see Section 1.2 of the online supplementary material) of the advisor and not only to accuracy per se: They relied more on advisors whose judgments were less redundant with their own.

On the contrary, when objective trial-by-trial feedback was removed, the pattern of results partly reversed such that competence ratings and influence were greatest for the advisor who tended to agree with participants' confidently made judgments. These advisors were more influential also on those occasions where they disagreed with participants' initial judgments, indicating that participants were actively learning about their advisors' overall competence, rather than simply discounting disagreeing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

advice (Section 2 of the online supplementary material). Surprisingly, participants did not perceive as least accurate the antibias advisor who agreed with them more frequently when their initial judgment was made with low confidence. If anything, trust was lower in the unbiased advisor. The difference was not reliable and hence must be interpreted with caution, but we note a possible link to an existing proposal suggesting that scarcely differentiated judgments across different observations tend to be indicative of lower expertise (Weiss & Shanteau, 2003). Importantly, regardless of the explanation of this unexpected result, the findings of Experiment 3 demonstrate participants' sensitivity to advisors in relation to their own confidence in their judgments and learn differentiated patterns of competence accordingly.

Network-Level Impacts of Individuals' Trust Strategies: An Agent-Based Simulation

The experimental and simulation results described above demonstrate that the absence of trial-level objective feedback does not preclude agents from inferring the competence of other agents. However, systematic deviations in perceptions of advisor competence can be observed between feedback and feedback-free scenarios when judges' and advisors' opinions are correlated (e.g., due to shared biases or common sources of information). To extend this work, we used agent-based modeling to explore how these effects might play out in more complex, multifactor situations.

If patterns of trust vary depending on feedback availability, we might expect different macrolevel patterns of trust in networks where feedback is available versus difficult to obtain. Furthermore, the presence of bidirectional information channels between agents (as opposed to judge-advisor systems) may lead to clustering of people sharing similar biases, increasing their polarization (i.e., confidence) and tendency to show herding behavior (Janis, 1972; Turner & Pratkanis, 1998). These effects parallel important observations about social networks, where echo-chambers and recommendation systems can produce clusters of individuals with similar characteristics that are impenetrable to external information (Del Vicario, Bessi, et al., 2016; Jasny et al., 2015; Pariser, 2011; Sunstein, 2001). Consuming within-cluster information more than between-cluster information can in turn lead to spurious consensus effect (Bessi, 2016; Del Vicario, Vivaldo, et al., 2016; Yaniv et al., 2009). Similarly, people in a group are known to have better access to shared (rather than private) information (Lightle, Kagel, & Arkes, 2009; Stasser & Titus, 2003). The present ideas provide a normative account of these otherwise suboptimal effects: According to our findings, in the absence of an objective standard, it may be adaptive to prefer advisors whose opinions agree with our own confidently held views. However, this strategy can lead to systematic biases in trust when opinions and judgments are nonindependent (a common feature of real social networks).

To explore the implications of these ideas, we manipulated agents' access to metacognitive signals in an agent-based simulation, and assessed how feedback availability interacts with cognitive mechanisms to produce emerging network macrostructures (Couzin, Krause, James, Ruxton, & Franks, 2002; Epstein, 2013). We expected that (a) when judgments in the population are independent, agreement-based heuristics are useful in reliably approximating true underlying expertise; (b) when judgments are correlated—for example, due to the presence of shared biases in the population—agents sharing similar biases will tend to cluster together due to the use of a confidence-agreement heuristic in feedback-poor (but not feedback-rich) scenarios; and (c) the use of internal representations of others' competence to discount advice will be beneficial to agents' performance when judgments are independent but not when they are correlated.

Model Description

Our agent-based models simulated the development of trust among agents as they make decisions and revise these decisions on the basis of learning the opinions of others. The decisions are simulated as a series of simple binary choices—is a stimulus from category A or B?—as a generalized case of the specific dot discrimination task used in our experiments. Of interest was how perceived competence ("trust") among agents was affected by the relative quality of their decisions, the availability of feedback and, crucially, the degree to which agents shared biases in their decision processes. Our models formalized these biases as differences in prior (base rate) expectations about the likelihood of A or B being the correct answer, which via Bayesian belief updating (Section 5 in the online supplementary material) will bias the interpretation of incoming information in the decision process (as an analogue of what might be expected to happen in everyday decisions as a function of our political leanings, news readings, musical preferences, etc.).

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, either without discounting advice or by discounting the advisor's judgment proportionally to their current level of trust in the advisor. After updating their judgments, each agent updated its trust judgments based on the available information about other agents (feedback or estimated partner's accuracy based on the agreement-in-confidence heuristics described above). We assessed the effect of feedback by parametrically manipulating its availability.

When learnt trust depends on agreement-based heuristics, levels of trust should track true accuracy when judgments are independent but generate clustering of populations—namely high within-group trust and low between-groups trust—when judgment correlations emerge within such populations. We test how network clustering is shaped by the presence or absence of objective feedback and show that bias-specific segregation arises only in the absence of feedback.

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 (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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

than expected, should reduce their bias toward *A*s. Conversely, bias should get stronger if the social contexts reinforces it (although see Bail et al., 2018; see the online supplementary material for details on network clustering computation and bias updating rules).

Results

We first tested the hypothesis that agreement-based heuristics are adaptive in situations of independent judgments because they reliably track others' expertise without the need to rely on external forms of feedback. To this end, we set all agents' initial bias $p(A)$ to 0.50 and the probability of feedback to 0. We then created two subpopulations of agents that orthogonally varied in their overall judgment accuracy (modeled as the degree of perceptual noise, with decreasing accuracy as a function of increasing noise) and calculated the average trust toward a target subpopulation (here, Population 2). The results show that average trust in Population 2 agents correctly tracks their underlying perceptual noise: Trust in Population 2 is highest when their perceptual noise is low and decreases as their noise increases (y-gradient in Figure 6A). Interestingly, the effect is not linear and we observe an interaction (x-gradient in Figure 6A) between the accuracy of Population 1 and the accuracy of Population 2 (no interaction is observed in Population 2's perception of itself, see Supplementary Figure S8 in the online supplementary material), mirroring the simple numerical analysis in Figure 1B: In the absence of feedback, poor performers in Population 1 fail to distinguish the accuracy of others, evident as a narrower range of trust values in rightward pixels in Figure 6A [compare (Kruger & Dunning, 1999)]. Nevertheless, overall, we see that trust in Population 2 systematically tracks its level of performance, even in the absence of objective feedback.

Actively inferring the competence of others is beneficial for agents. Population 1 agents generally improve their accuracy by receiving advice from Population 2 members (most pixels in Figure 6B are light-colored), but this benefit becomes an accuracy cost if Population 2 agents are much worse at the task than Population 1 agents (dark pixels in the lower left of Figure 6B; cf. Bahrami et al., 2010). However, this cost of receiving bad advice is mitigated when advice is discounted according to trust learnt through confidence-weighted agreement (Figure 6C, light-colored pixels in the lower left corner). Thus, learning differentiated patterns of trust enables agents to benefit from advice when it is useful and downweight or ignore it otherwise.

Our second hypothesis was that when the true state of the world is difficult to discern (e.g., when objective feedback is rare or absent, signal strength is weak or perceptual noise is large), bias-specific segregation can arise. We therefore ran a simulation in which we varied the bias $p(A)$ shown by two subpopulations of equal size and equal average accuracy (equivalent average perceptual noise). The two populations differed in their base rate estimates of the relative likelihood of the two outcomes (A and B), with Population 1 biased toward judging events as "B"s and Population 2 biased toward judging events as "A"s (i.e., $p(A)$ drawn in the range [0, .5] and [.5, 1], respectively). On each iteration, agents interacted with another randomly selected agent and updated their trust via a simple delta rule (Equation S25 in the online supplementary material). Figure 7 shows the average clustering—quantified as the degree to which agents learn greater trust in others who share their initial biases than others who have different initial biases—after 1,000 iterations. The data are plotted as a function of signal strength (low in Panel A to high in Panel B), perceptual noise (variation along the *x*-axis of each panel), and feedback probability (*y*-axis).

Average clustering decreased as the signal strength increased (Figure 7A–B panels), as perceptual noise increased (gradient over *x*-axis), and as the probability of receiving objective feedback increased (gradient over *y*-axis). In these cases, agents' decisions are dominated by external evidence rather than their prior expectations (Equation S20 in the online supplementary material) and within-group and between-groups agreement rates are similar. On the contrary, when feedback availability, noise or signal strengths decrease, decisions and thus agreement-based trust will more

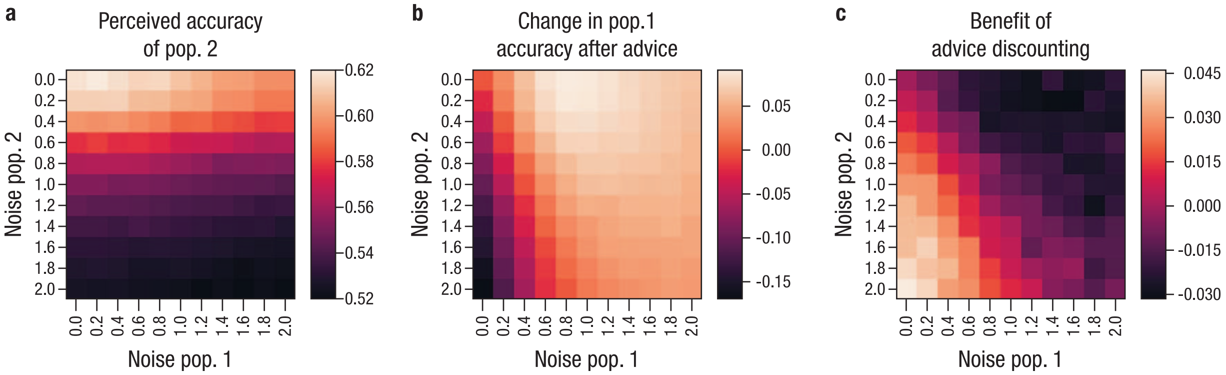

Figure 6. a: Average trust shown in agents belonging to Population 2. Trust in these agents increases as the noise affecting them decreases. Interestingly, trust formation is affected by perceptual noise of the observed agent(s) as well as perceptual noise of the observing agent(s) (compare Kruger & Dunning, 1999). b: Impact of advice for Population 1's accuracy. Pixels' color represent the difference between post and preadvice accuracy. c: The use of a trust-based advice discounting strategy increases postinteraction accuracy of accurate agents' but not inaccurate ones, particularly when interacting with low accuracy advisors (bottom left corner). See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

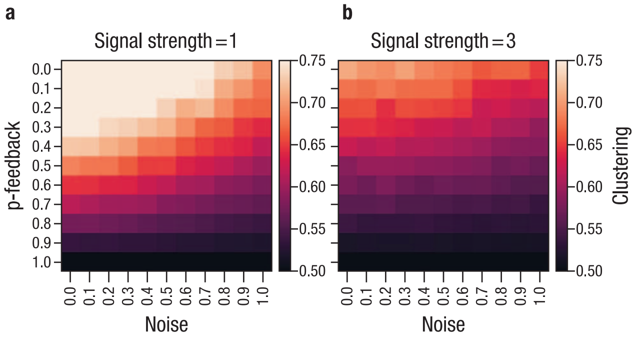

Figure 7. Clustering as a function of signal strength, probability of feedback and noise. Each pixel in each panel indicates the average clustering value (over 20 simulations), computed as described in the online supplementary material, after 1,000 iterations of decision, advice, and update in a network of agents with fixed stimulus strength, perceptual noise, and feedback probability parameters. A value of 0.5 indicates equal trust in in-group and out-group members. Values greater than 0.5 indicate greater trust in in-group than out-group members. The two panels represent increasing signal strengths. See the online article for the color version of this figure.

strongly be influenced by prior expectations, as predicted by Bayes's theorem. This finding suggests that simple rules of trust update can perform very differently depending on what learning signal is used. In the absence of objective feedback, the circular nature of using one's own belief to estimate others' accuracy produces higher clustering and segregation, particularly when decisions are often ambiguous. The presence of noise reduces the risk of getting stuck on local minima, particularly when correlations exist between judges (Couzin et al., 2011; Kao & Couzin, 2014; Shirado & Christakis, 2017).

Our third and final hypothesis was that segregation between in-group and out-group members, once established, can resist modification and in turn shape the evolution of shared beliefs. In this simulation, after 500 iterations we allowed agents to update their bias p(A) via delta rule, whereby their experience of the relative prevalence of A and B outcomes—as determined by their posterior opinion, opinions of partners, and objective feedback, when available—led to modification of their estimates of p(A). Of interest was the way in which the biases of the two groups evolved across interactions as a function of feedback availability, perceptual noise, and information sampling strategy.

When biased agents (differing in prior $p(A)$ and marked by different colors in Figure 8) select partners at random (columns A and B), or when objective feedback is available (saturated lines), decision biases rapidly diminish such that agents from both populations converge on accurate estimates of the relative likelihood of A versus B outcomes (i.e., $p(A) = 0.5$), independently of noise (Figure 8A and 8C). Columns B and D plot the distribution of biases across members of each group after the final iteration. However, when feedback is less available (low saturation lines) and trust drives agents' selection of their advisors (Figure 8C–8D), population biases are persistent in a manner that exhibits sensitivity to both information selection strategy and levels of perceptual noise. Cluster segregation is alleviated by greater perceptual noise (S9–S10) but magnified further if agents discount advice received based on trust (Supplementary Figure S11 in the online supplementary material). Thus, social information can stabilize (and even increase) initial group biases insofar as agents rely on this social information in lieu of objective feedback. Under these circumstances, a positive feedback loop can develop, whereby agents who share an initial bias (of, say, believing that p(A) > 0.5) will tend to ask each other for advice more often, thereby compounding each other's biased beliefs over time, thereby increasing their tendency to seek advice from like-minded agents, thereby compounding their bias, and so on across iterations.

In this way, our agent-based modeling simulations demonstrate how suboptimal behaviors at the network-level—such as self-sustaining biases and even extremes of polarization—can emerge from learning strategies that are normatively sensible at the level of individual agents: Using decision confidence as a proxy for objective accuracy can allow independent agents to learn about the validity of social information sources when feedback is infrequent or absent. Moreover, using social information can in turn allow agents to learn about the world under these conditions of diminished feedback. However, these learning strategies exhibit important limitations when agents' initial decisions are not independent but rather exhibit patterns of shared versus distinct biases across individuals within a population: They lead to predictable patterns of clustering and trust when agents' agreement rates are inflated by their shared biases or shared access to correlated information. Moreover, bias updating and partner-selection strategies also affect network structure by changing the segregation of individuals sharing the same bias, leading to stabilization of networks with systematically biased or even increasingly polarized beliefs about the world (in our simulations, the relative likelihood of A vs. B being the correct decision).

General Discussion

The present research explored how people learn the accuracy of others' advice in contexts where objective external feedback is not readily available or costly to acquire. In these scenarios, we hypothesize that metacognitive confidence provides a useful proxy for objective feedback. This is due to the fact that in many tasks, confidence provides a finely calibrated estimate of underlying accuracy (Fleming, 2016; Henmon, 1911; Koriat, 2012; Pescetelli et al., 2016). Thus, people can apply this internal confidence estimate to received advice in order to estimate its accuracy: as p(correct) in the case of agreement, and 1-p(correct) in the case of disagreement. This signal, accumulated over time, can support learning of reliable estimates of others' competence. Experiment 1 empirical data and models show the power of this agreement-in-confidence heuristic in identifying subtle differences in advisor quality when advisors provide new independent information: Participants (and simple models) learned to distinguish advisors of differing accuracy and confidence calibration even in the absence of objective feedback, and did so to a similar extent to participants (and simple models) who had access to objective feedback after every decision. Thus, we propose that people are able to follow internal signals (e.g., decision confidence) in a comparable manner to external signals—such as rewards or feedback—to learn about the accuracy of advice and advisors according to associative learning principles (Behrens et al., 2008; Sutton & Barto, 1998).

Experiments 2 and 3 tested a key prediction (and a crucial limitation of) the agreement-in-confidence heuristic, in common

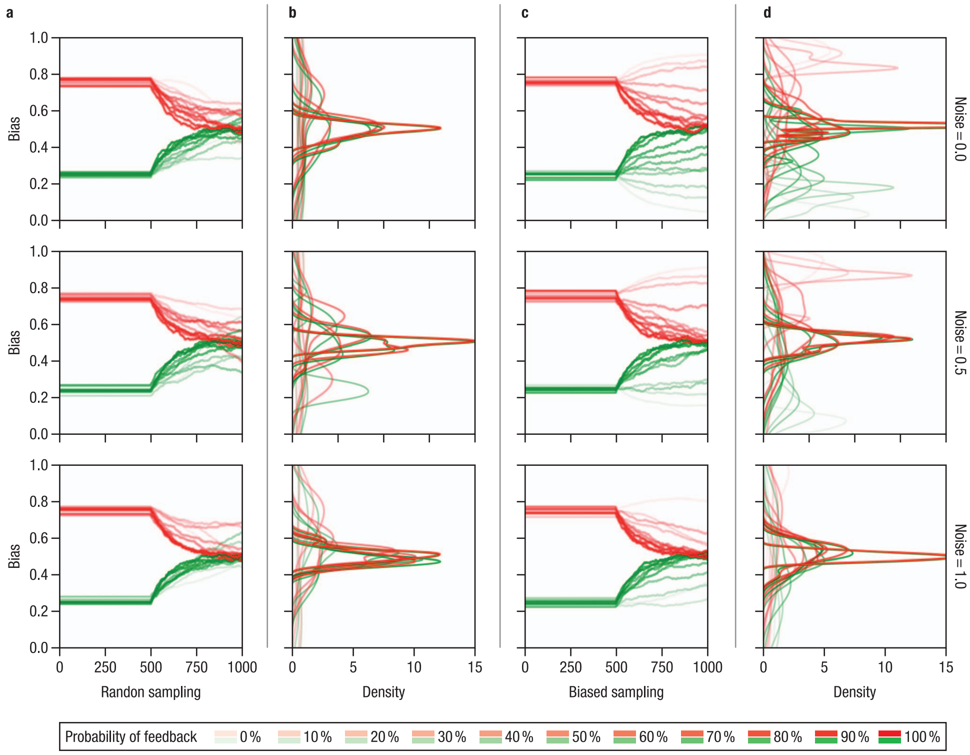

Figure 8. Bias as a function of time, feedback probability and agents' perceptual noise. Rows represent increasing levels of noise. Lines alpha values (saturation) represent probability of objective feedback. Columns a and b show agents' bias evolution over time and final bias distribution when agents choose their partners at random. Columns c and d show the same graphs when agents choose their partner proportionally to their trust. See the online article for the color version of this figure.

with an established research tradition exploring heuristics and biases in human judgments (Gigerenzer, 2008; Tversky & Kahneman, 1974): According to this tradition, our limited cognitive capacities mean that we must often adapt to use approximations or "short-cuts" to find good-enough solutions to otherwise intractable problems (Tversky & Kahneman, 1983). However, use of such heuristics can lead to systematic errors when their assumptions are violated. In the current work, the agreement-in-confidence heuristic provides a solution to the challenging problem of estimating the accuracy of information in the absence of an objective standard. This solution works well when agreement correlates with accuracy, as is the case where initial judgment and advice are independent (Experiment 1). However, the process goes astray when the independence between judgment and advice is broken, as in Experiments 2 and 3 where advice was contingent on participants' initial decision and expressed confidence. Adverse patterns of competence ratings and influence emerged when objective feedback was unavailable. Our findings are in agreement with studies showing that people often tend to ignore the correlation structure of advisors and heavily rely on agreement among information sources (Yaniv et al., 2009).

These findings extend our understanding of advice-taking and trust formation in three main ways. First, they extend our understanding of the uses of metacognitive signals in learning, by showing that learning takes place also in the absence of objective feedback or contextual cues. Previous studies have recognized the adaptive value of metacognitive monitoring in a range of cognitive tasks, like uncertainty monitoring (Yeung & Summerfield, 2012), information seeking (Desender, Boldt, & Yeung, 2018), and cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001). The present findings demonstrate that confidence is not only the end-product of information flow from the external stimulus to a perceptual inference, but it can feed back to help making inferences about external events, thus complementing objective learning signals (Behrens et al., 2008). Evidence seems to support this notion, as confidence signals have been shown to parallel reward prediction error signals in feedback-free situations (Guggenmos et al., 2016; Zylberberg et al., 2018).

Second, our findings extend understanding of metacognition in the social domain. Specifically, our results go beyond previous evidence that people downweight advice that disagrees with their initial opinion (Liberman et al., 2012; Yaniv, 2004), to explore how aggregated experiences of agreement and disagreement shape people's learned evaluations of overall advisor competence. Thus, our effects are observed in terms of patterns of trust and influence that reflect learning across repeated interactions, such that advisors who more regularly disagree with a participants' confidently made choices are less influential, even on those trials where their advice happens to agree with the participants' view. Conversely, we find that advisors who more regularly agree with the participant, particularly when the participant is confident in their initial choice, are more influential even on those trials where their opinion diverges

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

from the participants' (Section 2 in the online supplementary material).

A long tradition in social psychology has investigated the importance of agreement and confidence in advice taking (Bonaccio & Dalal, 2006; Ecken & Pibernik, 2016; Swol & Sniezek, 2005), persuasion (Price & Stone, 2004), influence (Liberman et al., 2012; Rader et al., 2017; Sah et al., 2013), and group dynamics (Sherif et al., 1965; Stasser & Davis, 1981), and its relevance in applied fields such as judicial systems and organizations (Bovens & Hartmann, 2004; Roediger III et al., 2012; Schum & Martin, 1982; Stasser & Davis, 1981). This literature consistently demonstrates that people tend to pay an accuracy cost for not using advice enough or using it in a self-serving manner (Minson et al., 2011; Soll & Mannes, 2011); for example, by discounting advice proportionally to the conflict with one's own opinion. Here, we suggest that these phenomena might emerge as by-products of a normatively justified heuristic that tries to overcome the computational intractability of learning without objective feedback by approximating it with subjective probabilistic estimates.

Our third and final contribution is in identifying the boundary conditions of this best-response strategy. Other studies in the social domain have focused on the way that advisor's confidence provides a useful signal that can benefit group decision making (Bahrami et al., 2010; Koriat, 2012; Sorkin, Hays, & West, 2001) and social coordination (Shea et al., 2014). Most previous literature has focused on situations where observers are independent. Here, we investigate environments characterized by correlated information among judges, where the heuristics that people use to gauge the usefulness of social signals turn out to be maladaptive. Far from being fringe cases, these environments may be common in many real world scenarios, both in physical and digital space (Del Vicario, Vivaldo, et al., 2016; Kao et al., 2014). Correlation among judges emerge from sharing similar social or information cliques, which in turn lead to sharing similar biases. Our findings (Experiments 2–3, and agent-based simulation) show the micro- and macroscale effects of using confidence and agreement in feedback-poor information environments.

Our results were generally in line with the idea that we trust advisors according to their objective accuracy when feedback is present, and trust agreeing-in-confidence advisors when it is absent. However, there were some notable exceptions. First, advisor agreement rate had an effect over and above accuracy even when feedback was available on every trial in Experiment 2, suggesting that people value agreement even when redundant. Second, people seem to value information over pure accuracy, as shown by participants in the Feedback group of Experiment 3, who preferred the advisor who agreed with them more frequently when they themselves were correct but low in confidence (the antibias advisor) over equally accurate but less informative advisors. Finally, participants in Experiment 3 did not prefer the neutral to antibias advisor when feedback was absent. We explain this result in light of the fact that, without feedback, expertise can still be estimated by comparing the classification variability observed for similar stimuli with the classification variability observed across different stimuli (Weiss & Shanteau, 2003).

Our agent-based models allowed us to explore how these cognitive mechanisms might scale up in larger population interactions. In the models, we observed that when feedback was unavailable, and agents' judgments were independent, agreement-based trust formation strategies helped agents trust more accurate advisors. However, when individuals covaried in their signals (e.g., by sharing systematic biases), individuals segregated according to their initial biases. Agents within a homogeneous population were more likely to trust and influence each other than agents belonging to different populations. The polarization of each cluster's average bias was reduced by the presence of random noise, increased signal strength or the presence of objective feedback.

These results provide a potential new understanding of echo-chambers and assortative mixing online (Bollen, Gonçalves, Ruan, & Mao, 2011; Sunstein, 2001), as a byproduct of an otherwise adaptive mechanism—use of an agreement-in-confidence heuristic to estimate advice accuracy—when it is difficult to objectively assess the validity of others' opinions, and when these opinions might themselves be subject to systematic shared biases. In these scenarios, random fluctuations in initial bias for one option within a population of equally accurate individuals can spiral out to form densely connected communities, thus effectively modifying the network structure. The results show how simulated agents endowed with realistic cognitive mechanisms can shed light on the emergence of complex patterns at the population level (Epstein, 2013). More broadly, the present research suggests the value of identifying strategies used by individual decision makers—who are likely to rely on imperfect heuristics given necessary limits in the information they access and the cognitive resources they have available to process that information—and exploring how these strategies can influence behavior at the group and network level. The debate around the false consensus effect (Dawes, 1989; Krueger & Clement, 1994; Ross et al., 1977) is a clear example of how understanding biases in terms of their adaptive potential can foster fruitful debate in psychology. In a similar vein, we hope our work can further understanding of phenomena like echo-chambers and assortative mixing, to design better platforms for democratic debate that are mindful of the heuristics they might elicit in their users.

Conclusions

The current work aimed to show that confidence is a valuable attribute of someone's judgment in social decision-making. It helps others discriminate when one is more likely to be correct (and thus value their contribution) but also helps a decision-maker to make consistent judgments about others irrespective of feedback availability. However, this potentially adaptive solution to an intractable problem of learning in the absence of feedback can backfire when judgments from different observers are not independent, leading to systematic biases in trust and influence.

Context of the Research

This research was conducted as part of the D.Phil. research of Niccolò Pescetelli. It was inspired by research on the computational mechanisms of social interaction, for example, by Bahador Bahrami (with whom Niccolò Pescetelli completed his MSc dissertation), and the confluence of these ideas with research in Nicholas Yeung's lab concerning the nature and functional role of confidence judgments. Although confidence has for a long time been studied in terms of its role in social and organizational decision making, new developments in cognitive neuroscience

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

place metacognitive signals within the wider context of cognitive control and uncertainty monitoring. Working within this framework, the present research adds to growing evidence—from Nicholas Yeung's lab and elsewhere—indicating that metacognitive signals play an important role in adaptive behavior, both at the level of individual decision makers and at the level of groups of interacting individuals. This research also sits well within the research program of Niccolò Pescetelli, whose research focuses on social learning and decision-making by human and algorithmic agents. This work is being developed, although in different directions, by both authors in their respective groups: the ACC lab, led by Nicholas Yeung at the University of Oxford, and the Hybrid Collective Intelligence group, led by Niccolò Pescetelli at the Max Planck Institute for Human Development. The former is investigating how decision confidence drives the collection of new evidence, while the latter develops the current findings to hybrid human-machine collective decisions.

References

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology, 11,* 1.

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron, 81,* 195–206. http://dx.doi.org/10.1016/j.neuron.2013.10.018

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329,* 1081–1085. http://dx.doi.org/10.1126/science.1185718

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

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37,* 379–384.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456,* 245–249. http://dx.doi.org/10.1038/nature07538

Bessi, A. (2016). Personality traits and echo chambers on Facebook. *Computers in Human Behavior, 65,* 319–324. http://dx.doi.org/10.1016/j.chb.2016.08.016

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience, 35,* 3478–3484. http://dx.doi.org/10.1523/jneurosci.0797-14.2015

Bollen, J., Gonçalves, B., Ruan, G., & Mao, H. (2011). Happiness is assortative in online social networks. *Artificial Life, 17,* 237–251. http://dx.doi.org/10.1162/artl_a_00034

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101,* 127–151. http://dx.doi.org/10.1016/j.obhdp.2006.07.001

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652. http://dx.doi.org/10.1037/0033-295X.108.3.624

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. Oxford, UK: Oxford University Press.

Clark, C. C., & Lawless, H. T. (1994). Limiting response alternatives in time-intensity scaling: An examination of the halo-dumping effect. *Chemical Senses, 19,* 583–594. Retrieved from https://psycnet.apa.org/record/1995-27840-001

Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., . . . Leonard, N. E. (2011). Uninformed individuals promote democratic consensus in animal groups. *Science, 334,* 1578–1580. http://dx.doi.org/10.1126/science.1210280

Couzin, I. D., Krause, J., James, R., Ruxton, G. D., & Franks, N. R. (2002). Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology, 218,* 1–11. http://dx.doi.org/10.1006/jtbi.2002.3065

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25,* 1–17. http://dx.doi.org/10.1016/0022-1031(89)90036-X

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 554–559. http://dx.doi.org/10.1073/pnas.1517441113

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on Facebook. *Scientific Reports, 6,* 37825. http://dx.doi.org/10.1038/srep37825

De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., & Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron, 79,* 1222–1231. http://dx.doi.org/10.1016/j.neuron.2013.07.003

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29,* 761–778. http://dx.doi.org/10.1177/0956797617744771

Ecken, P., & Pibernik, R. (2016). Hit or miss: What leads experts to take advice for long-term judgments? *Management Science, 62,* 2002–2021. http://dx.doi.org/10.1287/mnsc.2015.2219

Epstein, J. M. (2013). *Agent_zero: Toward neurocognitive foundations for generative social science*. Princeton, NJ: Princeton University Press.

Fleming, S. M. (2016). Changing our minds about changes of mind. *eLife*. Advance online publication. http://dx.doi.org/10.7554/eLife.14790

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision performance: A general Bayesian framework for metacognitive computation. *Psychological Review, 124,* 1–59. http://dx.doi.org/10.1037/rev0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*. Advance online publication. http://dx.doi.org/10.3389/fnhum.2014.00443

Gigerenzer, G. (2008). Why heuristics work? *Perspectives on Psychological Science, 3,* 20–29.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Guggenmos, M., & Sterzer, P. (2017). A confidence-based reinforcement learning model for perceptual learning. *bioRXiv*. http://dx.doi.org/10.1101/136903

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*. Advance online publication. http://dx.doi.org/10.7554/eLife.13388

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review, 18,* 186–201. http://dx.doi.org/10.1037/h0074579

Hertz, U., Romand-Monnier, M., Kyriakopoulou, K., & Bahrami, B. (2016). Social influence protects collective decision making from equality bias. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 164–172. http://dx.doi.org/10.1037/xhp0000145

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford, UK: Oxford University Press. Retrieved from https://books.google.co.uk/books?id=139Oa4MOsAgC&redir_esc=y

Janis, I. L. (1972). *Victims of groupthink*. Boston, MA: Houghton Mifflin.

Jasny, L., Waggle, J., & Fisher, D. R. (2015). An empirical examination of echo chambers in U.S. climate policy networks. *Nature Climate Change, 5,* 782–786. http://dx.doi.org/10.1038/nclimate2666

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Royal Society B: Biological Sciences, 281, 20133305. http://dx.doi.org/10.1098/rspb.2013.3305

Kao, A. B., Miller, N., Torney, C., Hartnett, A., & Couzin, I. D. (2014). Collective learning and optimal consensus decisions in social animal groups. *PLoS Computational Biology, 10*(8), e1003762. http://dx.doi.org/10.1371/journal.pcbi.1003762

Koriat, A. (2012). When are two heads better than one and why? *Science, 336,* 360–362. http://dx.doi.org/10.1126/science.1216549

Krause, J., Ruxton, G. D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology & Evolution, 25,* 28–34. http://dx.doi.org/10.1016/j.tree.2009.06.016

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67,* 596–610. http://dx.doi.org/10.1037/0022-3514.67.4.596

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,* 1121–1134.

Le Bon, G. (1895). *La psychologie des foules* [Crowd psychology] (Félix Alca, Ed.). Paris, France: Ancienne Librairie Germer Bailliere et Cie. Retrieved from http://socserv.mcmaster.ca/econ/ugcm/3ll3/lebon/Crowds.pdf

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads". *Journal of Experimental Social Psychology, 48,* 507–512. http://dx.doi.org/10.1016/j.jesp.2011.10.016

Lightle, J. P., Kagel, J. H., & Arkes, H. R. (2009). Information exchange in group decision making: The hidden profile problem reconsidered. *Management Science, 55,* 568–581. http://dx.doi.org/10.1287/mnsc.1080.0975

Mackay, C. (1841). *Extraordinary popular delusions and the madness of crowds*. Ware, UK: Wordsworth Edition Limited.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., . . . Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 3835–3840. http://dx.doi.org/10.1073/pnas.1421692112

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20,* 709. http://dx.doi.org/10.2307/258792

Meyniel, F., Sigman, M., & Mainen, Z. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron, 88,* 78–92. http://dx.doi.org/10.1016/j.neuron.2015.09.039

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango. *Personality and Social Psychology Bulletin, 37,* 1325–1338. http://dx.doi.org/10.1177/0146167211410436

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83,* 602–627. http://dx.doi.org/10.1037/0033-2909.83.4.602

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. London, UK: Penguin.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1,* 817–845.

Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2020). *Confidence drives post-decisional search of information from social sources*. Manuscript in preparation.

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General, 145,* 949–965. http://dx.doi.org/10.1037/xge0000180

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience, 19,* 366–374. http://dx.doi.org/10.1038/nn.4240

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17,* 39–57. http://dx.doi.org/10.1002/bdm.460

Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General, 147,* 1431–1444. http://dx.doi.org/10.1037/xge0000471

Rader, C. A., Larrick, R. P., & Soll, J. B. (2017). Advice as a form of social influence: Informational motives and the consequences for accuracy. *Social and Personality Psychology Compass, 11*(8), e12329. http://dx.doi.org/10.1111/spc3.12329

Roediger, H. L., III, Wixted, J. H., & Desoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. P. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–117). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199920754.003.0004

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13,* 279–301. http://dx.doi.org/10.1016/0022-1031(77)90049-X

Sah, S., Moore, D. A., & Maccoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes, 121,* 246–255. http://dx.doi.org/10.1016/j.obhdp.2013.02.001

Schultze, T., Gerlach, T. M., & Rittich, J. C. (2018). Some people heed advice less than others: Agency (but not communion) predicts advice taking. *Journal of Behavioral Decision Making, 31,* 430–445. http://dx.doi.org/10.1002/bdm.2065

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law & Society Review, 17,* 105–152. http://dx.doi.org/10.2307/3053534

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18,* 186–193. http://dx.doi.org/10.1016/j.tics.2014.01.006

Sherif, C., Sherif, M., & Nebergall, R. (1965). *Attitude and attitude change*. Philadelphia, PA: W. B. Saunders Company.

Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature, 545,* 370–374. http://dx.doi.org/10.1038/nature22332

Shultz, T. R., Katz, J. A, & Lepper, M. R. (2001). Clinging to beliefs: A constraint-satisfaction model. *Proceedings of the Annual Meeting of the Cognitive Science Society*. Retrieved from https://escholarship.org/uc/item/0qh094sh

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization, 1,* 161–176.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes, 62,* 159–174. http://dx.doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes, 84,* 288–307. http://dx.doi.org/10.1006/obhd.2000.2926

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 780–805. http://dx.doi.org/10.1037/a0015145

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting, 27,* 81–102. http://dx.doi.org/10.1016/j.ijforecast.2010.05.003

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review, 108,* 183–203. http://dx.doi.org/10.1037/0033-295X.108.1.183

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Stasser, G., & Davis, J. H. (1981). Group decision making and social influence: A social interaction sequence model. *Psychological Review, 88,* 523–551. http://dx.doi.org/10.1037/0033-295X.88.6.523

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry, 14,* 304–313.

Sunstein, C. R. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology, 44,* 443–461.

Tenney, E. R., MacCoun, R. J., Spellman, B. a., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18,* 46–50. http://dx.doi.org/10.1111/j.1467-9280.2007.01847.x

Tost, L. P., Gino, F., & Larrick, R. P. (2012). Power, competitiveness, and advice taking: Why the powerful don't listen. *Organizational Behavior and Human Decision Processes, 117,* 53–65. http://dx.doi.org/10.1016/j.obhdp.2011.10.001

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35,* 2503–2522.

Turner, M. E., & Pratkanis, A. R. (1998). Twenty-five years of groupthink theory and research: lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes, 73,* 105–115. http://dx.doi.org/10.1006/obhd.1998.2756

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. Retrieved from https://science.sciencemag.org/content/185/4157/1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315. http://dx.doi.org/10.1037/0033-295X.90.4.293

Vandormael, H., Herce Castañón, S., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences of the United States of America, 114,* 2771–2776. http://dx.doi.org/10.1073/pnas.1613950114

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human factors, 45,* 104–116. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12916584.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93,* 1–13. http://dx.doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology Learning, Memory, and Cognition, 35,* 558–563. http://dx.doi.org/10.1037/a0014589

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83,* 260–281. http://dx.doi.org/10.1006/obhd.2000.2909

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical transactions of the Royal Society of London*. Series B, Biological Sciences, *367,* 1310–1321. http://dx.doi.org/10.1098/rstb.2011.0416

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33,* 345–366. http://dx.doi.org/10.1006/jesp.1997.1326

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron, 99,* 1083–1097. http://dx.doi.org/10.1016/j.neuron.2018.07.035

(*Appendix follows*)

Appendix

Summary Tables of Main Across-Experiments Effects

Table A1
Across-Experiment Main Effects on Rated Competence as Reported in the Main Text

	Rated competence		
Factor	$F(1, 44)$	p value	η^2_G (ϵ)
Experiment 1			
Fb	$F(1, 44) < 1$	$p > .99$	.00
Acc	$F(1, 44) = 9.68$	$p = .003^{**}$	.079
Cal	$F(1, 44) = 12.32$	$p = .001^{**}$	.076
Fb:Acc	$F(1, 44) < 1.9$	$p > .16$	.01
Fb:Cal	$F(1, 44) < 1.9$	$p > .16$	.01
Acc:Cal	$F(1, 44) < 1.9$	$p > .16$	.01
Fb:Acc:Cal	$F(1, 44) < 1.9$	$p > .16$	.01
Experiment 2			
Fb	$F(1, 44) < 1$	$p > .45$	2.19e-04
Acc	$F(1, 44) = 8.36$	$p = .005^{**}$	.06
Agr	$F(1, 44) = 22.52$	$p < .001^{**}$	.1
Fb:Acc	$F(1, 44) = 8.41$	$p = .005^{**}$	.06
Fb:Agr	$F(1, 44) = 1.88$	$p = .17$	.1
Acc:Agr	$F(1, 44) < 1$	$p > .83$	2.03e-05
Fb:Acc:Agr	$F(1, 44) < 1$	$p > .84$	3.04e-04
Experiment 3			
Fb	$F(1, 44) < 1$	$p > .99$	1.39e-33
Typ	$F(2, 92) = 1.66$	$p = .19$	.03($\epsilon = .99$)
Fb:Typ	$F(1, 92) = 6.64$	$p = .002^{**}$	.12($\epsilon = .99$)

Note. ϵ = Greenhouse-Geisser sphericity correction parameter; Fb = feedback; Acc = accuracy; Cal = calibration; Agr = agreement; Typ = advisor type.
** $p < .01$.

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table A2

Across-Experiment Main Effects on Influence as Reported in the Main Text

	Influence		
Factor	*F(1, 44)*	*p* value	η^2_G (ϵ)
Experiment 1			
Fb	0.02	8.73e-01	3.98e-04
Acc	14.80	3.81e-04**	2.36e-02
Cal	15.84	2.54e-04**	1.38e-02
AC	55.82	2.34e-09**	1.22e-01
Fb:Acc	0.04	8.40e-01	6.67e-05
Fb:Cal	0.60	4.42e-01	5.32e-04
Fb:AC	1.16	2.85e-01	2.92e-03
Acc:Cal	0.24	6.20e-01	2.83e-04
Acc:AC	2.73	1.05e-01	1.10e-03
Fb:Acc:Cal	0.80	3.73e-01	9.19e-04
Fb:Acc:AC	0.65	4.21e-01	2.67e-04
Fb:Cal:AC	0.93	3.37e-01	3.97e-04
Acc:CalAC	4.75	3.45e-02**	8.58e-04
Fb:Acc:Cal:AC	3.92	5.38e-02	7.08e-04
Experiment 2			
Fb	3.19	.08	4.68e-02
Acc	14.79	.0003***	4.32e-02
Agr	13.91	.0005***	3.37e-02
Fb:Acc	0.56	.45	1.73e-03
Fb:Agr	2.01	.16	5.03e-03
Acc:Agr	0.01	.91	1.97e-05
Fb:Acc:Agr	0.26	.61	4.66e-04
Experiment 3			
Fb	$F(1, 46) = 1.36$	.24	0.017
Typ	$F(2, 92) = 1.12$	.33	0.009 ($\epsilon = .77$)
Fb:Typ	$F(2, 92) = 4.80$	.01*	0.039 ($\epsilon = .77$)

Note. ϵ = Greenhouse-Geisser sphericity correction parameter; Fb = feedback; Acc = accuracy; Cal = calibration; Agr = agreement; Typ = advisor type. Unless otherwise specified, degrees of freedom for the *F* statistic are (1,44).

* $p < .05$. ** $p < .01$. *** $p < .001$.

Received November 1, 2019
Revision received June 24, 2020
Accepted June 29, 2020 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.