Do humans recalibrate the confidence of advisers or take confidence at face value?

Oana Stanciu (stanciuo@phd.ceu.edu)

Jozsef Fiser (fiserj@ceu.edu)

Department of Cognitive Science, CEU,
Quellenstraße 51, Vienna, Austria

Abstract

Who we choose to learn from is influenced by the relative confidence of potential informants (Birch, Akmal, & Frampton, 2010). More confident advisers are preferred based on an assumption that confidence is a good indicator of accuracy. However, oftentimes, accuracy and confidence are not calibrated, either due to strategic manipulations of confidence or unintentional failures of metacognition. When accuracy information is readily available, people are additionally vigilant to the calibration of informants, penalizing incorrect, yet confident advisers (Tenney, MacCoun, Spellman, & Hastie, 2007). The current experiment tested whether participants can leverage inferences about two advisers' calibration profiles to make optimal trial-by-trial decisions. We predicted that choice of advisers reflects relative differences in the advisers' probability of being correct given their stated confidence (recalibrated confidence), as opposed to stated confidence differences. The prediction was not supported by data, but calibration had a modulating effect on choices, as more confident advisers were more influential only when they were also calibrated. Further, participants' decision confidence was informed only by the confidence of the adviser whose advice was chosen, disregarding the confidence of the second adviser.

Keywords: metacognition; overconfidence; calibration; adviser preference

Introduction

Human reliance on social learning makes us susceptible to being misinformed by others, and, therefore, provides an incentive for the development of epistemic vigilance (Sperber et al., 2010). This is particularly salient in the case of expressions of confidence as strategic manipulations of confidence are an effective means of gaining influence (Kurvers et al., 2021). Tracking the metacognitive states of others is also relevant in cooperative situations, as there is wide (and consistent) inter-individual variability in the metacognitive performance (Song et al., 2011).

In the absence of external evidence, people employ a confidence heuristic (Thomas & McFadyen, 1995), assuming that a more confident agent is also more accurate, which should lead to efficient information exchange if agents exhibit perfect metacognitive sensitivity. Previous judge-adviser experiments (Price & Stone, 2004) have shown that participants employ the confidence heuristic even when assessing overconfident forecasters. On the other hand, when clear evidence is provided that directly contradicts the statements made by a confident individual, they are penalized. For instance, in a vignette study using the high stakes situation of a court trial, Tenney et al. (2007) found an interaction between the accuracy and confidence of eyewitnesses. Participants perceived highly confident witnesses as more credible than unconfident ones when they were correct (or in the absence of knowledge about the veridity of their testimony), while the opposite pattern was observed for inaccurate eyewitnesses.

In a study that required participants to make repeated decisions alone and then revise them following adviser recommendations, Sah, Moore, and MacCoun (2013) also found evidence for a default confidence preference, but it could be easily overturned by brief objective feedback about the adviser's performance at the beginning of the task. As predicted, participants' explicit ratings of the credibility of advisers were influenced by the calibration of advisers. On the other hand, Sah et al. (2013) did not find an effect of the calibration profile on how much the advice modified, trial-by-trial, the original judgements of participants, even when performance-based monetary incentives were added. We believe the null results may be explained by the fact that, intuitively, as the advisers presented were always correct or incorrect, there was no benefit (for improving participant estimation) in tracking confidence or modulating decisions based on confidence in a trial-by-trial fashion. The more disquieting alternative explanation is that miscalibration affects the credibility of informants as explicitly reported (as it may be interpreted as a sign of bad faith), without any implicit consequences for behavior in terms of discounting their advice in future interactions.

A line of studies in which the precise quantification of metacognitive abilities was possible (although not the primary focus) also suggests humans exhibit little vigilance towards other's expressions of confidence. Bahrami et al. (2010) showed that collaborators on a perceptual decision making task, if communicating freely, weight their individual decisions in proportion to their confidence to make joint decisions (that are generally superior to individual ones). Bahrami et al. (2010) assumed that participants were faithfully communicating confidence in accordance with their internal model. However, subsequent work has found that people exhibit an equality bias, downplaying differences in the reliability of collaborators (Mahmoodi et al., 2015), and wrongly assume that collaborators have equal metacognitive sensitivity (Pescetelli, Rees, & Bahrami, 2016). It is unclear whether participants would generally apply these assumptions beyond perceptual decision making, in more ab-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

stract tasks (where these assumptions are less warranted, (e.g. (Martí, Mollica, Piantadosi, & Kidd, 2018)). More importantly, given known self-placement biases, it would be relevant to explore whether participants behave differently when comparing the metacognitive skills of two informants as opposed to relating their own metacognitive ability to that of a collaborator in an interactive effort to solve a joint task.

To sum up, the relationship between one's confidence and accuracy certainly influences their perceived credibility and sway on others. However, based on the literature presented thus far, it is not clear how refined this ability is beyond cases of flagrant misrepresentation of one's knowledge. In the first set of studies discussed above, the advisers judged by participants were either entirely accurate or inaccurate (in sometimes only one-shot advice), and their confidence was also discretized or at least highly contrasting. Further, effects were only observed on overall preferences for an adviser, and not in the quantitative influence they exerted. In the latter set of studies, tracking metacognitive sensitivity of another agent was very difficult, and the joint nature of the task may have lead to equality biases.

A substantive test of metacognitive monitoring should involve testing the extent to which humans make optimal adviser choices given the relationship between accuracy and confidence. Specifically, whether people leverage the functional mapping between accuracy and confidence in order to determine, on every given encounter, the probability of advisers being correct given their stated confidence (and any other potential contextual information). We refer to this quantity as recalibrated confidence, in contrast to the explicitly stated confidence of advisers. The explicit and recalibrated confidence are the same only for calibrated advisers.

We propose a simple experiment in the judge-adviser framework in which participants could infer the relationship between the accuracy and confidence of two agents (manipulated across conditions) by observing them repeatedly perform a novel task. Following this, participants made multiple decisions relying solely on disagreeing advice from the potential advisers. We hypothesized that trial-by-trial, participants will choose the suggestion of the adviser with the highest recalibrated confidence as opposed to the highest stated confidence. In the current experiment, the optimal recalibration strategy leads to sometimes selecting the advice of an adviser who is explicitly less confident than their competitor independent of calibration. Thus, current predictions sometimes disagree with both the confidence heuristic and the calibration hypothesis.

Task intuition

Imagine you are a student struggling with solving an equation. You have two classmates, Anna and Emma, who both scored 70% in the latest math test. Anna says she is 90% confident she can solve your equation, while Emma rates her chances at 70%. They did equally well on the test, so who do you ask for help? At face value, given an assumption that accuracy and confidence go hand in hand, you should ask Anna to help. However, if you also know that after the last math test Anna thought she was 100% correct and Emma's confidence was at 60%, you can factor in Anna's overestimation of confidence and Emma's slight underestimation. Chances of Anna getting it right are likely around 60% and Emma's around 80% so in the end you are likely better off asking Emma for help.

Methods

Participants 60 participants, half female and equally split in the two between participant conditions, were recruited online through the Prolific platform. The mean age of participants was 33.27 years (range: 18 - 66 years old). English was the first language of all participants. Participation was rewarded with £6.5, this included a participation reward of £5.5 and a performance-based bonus of £1.

Task First, participants observed a pair of agents performing a simple task across several trials. Participants then engaged in a betting task in which they had to make decisions based on the advice of these agents. Lastly, participants answered questions about the performance of the two observed agents and stated their overall preference for one of them.

During the **observation phase**, participants saw two other fictitious participants get tested on a binary categorization task while stating their confidence in the correctness of their response on a continuous scale. Specifically, the agents were betting one virtual coin every trial on whether a "Modi alien" was depicted in the image on the left or on the right side of the screen. It was stressed that these two agents were not communicating and could not see each other, but were performing the task individually on the Prolific platform.

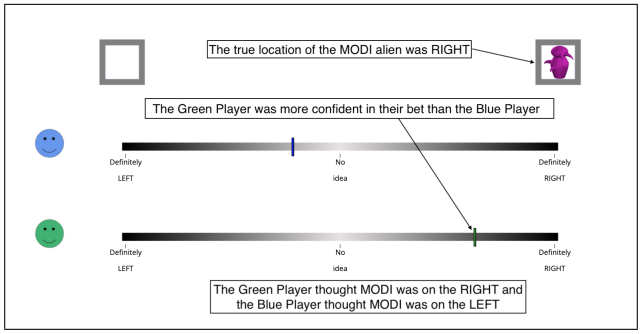

Figure 1: Example trial from the observation phase.

The agents expressed their answer and confidence simultaneously: the direction of the slider position relative to the center of the scale marked their decision (the alien is left or right) and the distance from the center of the slider marked how confident the participant was (see Figure 1). Only verbal labels were presented on the scale such that the center was "I don't know" and the extremes were marked with "Definitely right" or "Definitely left". Further, a grayscale gradient was used to mark the increase in confidence. Following the agents' decisions, the true location of the alien was presented

on the screen. Prior to the start of the experiment, participants were instructed to interpret values on the gradient scale as proportional to the estimated probability of being correct.

The observation phase consisted of 120 trials. Randomly interleaved attention checks were presented following 10% of trials to ensure task compliance. Participants were asked to make a two alternative forced choice (2AFC) about whether a given agent was correct in their answer in the previous trial. Feedback was provided.

Every 30 trials, participants were shown the number of correct answers and the average confidence so far for the two agents. Both summary statistics were presented on verbally labelled continuous scales. This means that participants were reminded of the summary statistics at the end of the observation phase.

The agents' decisions were presented alone on the screen for 2,500ms, the true location was then added and presented on screen for 2,000ms and the intertrial interval was 1,500ms. Answers to the attention checks were not speeded, and the feedback was presented for 500ms. Participants were allowed to consult the summary boards for as along as they wanted.

In the **betting phase**, participants were instructed that it was their turn to perform the categorization task. To incentivize performance, participants were told that on each trial they would have to bet one virtual coin on their answer, and as a function of that number of coins they won by the end of the task, they could receive a monetary bonus. Participants were only informed how many coins they earned at the end of the experiment. In total, participants made 60 bets.

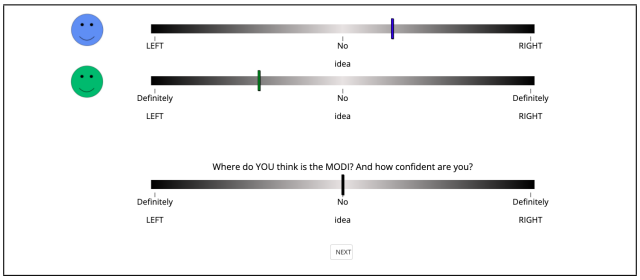

Figure 2: Example trial from the betting phase. Here, the participant chose the advice of the agent represented by the green avatar.

As before, participants could see the decisions and associated confidence of the two agents from the observation phase. Importantly, participants were not shown the two images in which the "Modi alien" could be depicted so they had no way of knowing the location of the alien by themselves. Thus, to make a 'blind' forced decision, participants could only rely on the decision made by the two agents and their associated decision confidence. Participants responses were made on a scale identical to the one used by the two agents. As such, participants simultaneously made a decision and expressed their level of confidence in that decision. Answers were unspeeded and no feedback was provided.

In the **post-test phase**, participants were asked to estimate the accuracy and confidence of the two agents on continuous scale. Participants then made a 2AFC decision on partner preference for a similar future task.

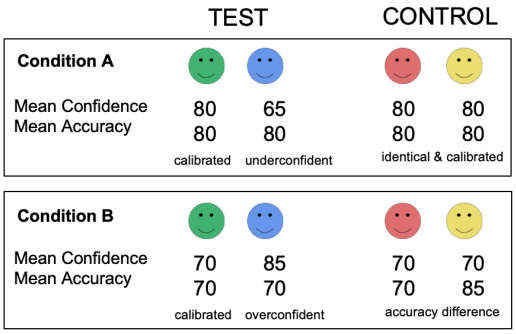

Figure 3: Task design. Test and Control blocks were run within-participant. Conditions A and B were between participant conditions. Numerical differences presented here were linearly scaled for visual presentation in the experiment (50% confidence corresponding to the center of the scale, and 100% to the extremes).

Design Each participant completed the task described above twice, with two different pairs of (fictional) advisers, the order of which was counterbalanced. In addition to this within-participant manipulation, there was also a between participant manipulation of the agent pairings. Figure 3 illustrates the design of the experiment.

We refer to one of the within-participant tasks as the Test block and one as the Control block. Control blocks were used as sanity checks of the design, and provided a measure of the minimum and maximum effects that can be expected. The control block of condition A used two agents matched for accuracy and confidence. The same agent was presented twice, with shuffled trial order. Any differences observed in this block can only be attributed to perceptual noise. In the control block of condition B, the two agents had the same confidence, but one of them was more accurate. This modulation should induce strong preferences, so a failure of to elicit a statistically significant effect would mean that participants did not learn the statistics of the task.

In in the test block of condition A, participants observed an underconfident agent paired with a calibrated one. In the test block of condition B, an overconfident agent was displayed alongside a calibrated one. Importantly, in both test blocks, the two agents had the same overall accuracy and the same approximately linear relationship between accuracy and confidence (see Figure 4). In addition, the standard deviation of the confidence ratings of the two agents was also approximately matched ($SD = 10$).

Since the accuracy of the agents was on individual trials was pseudo-randomized, the rate of disagreement varied among participants (with a mean of roughly 60% and 70%

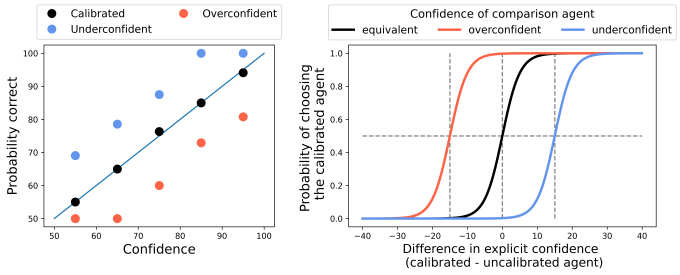

Figure 4: Left: Relationship between confidence and probability of being correct for agents with the same accuracy, but different marginal confidence. Right: Predictions for binary adviser choices as a function of confidence differences between advisers.

for the conditions where agents had 70% and 80% accuracy). Importantly, when the agents disagreed in the test conditions, they were equally likely to be correct. The correlation of confidence judgements of the two agents varied between participants, but was low ($r < .3$).

The confidence values of the agents in betting phase trials were selected to make it possible to distinguish whether participants chose trial-by-trial the agent with the highest recalibrated or explicit confidence. First, in all betting trials, the two agents disagreed about the location of the alien. Participants were explicitly informed that only trials with disagreements will be shown, since their decision would be obvious when the two agents agree. Second, we ensured that there were sufficient numbers of trials (33%) in which decisions based on explicit confidence differences were different than decisions based on recalibrated confidence, but shy of making decisions based on recalibration overwhelmingly favour one agent. Lastly, the average confidence of the agents was roughly equal in the betting trials.

The choices made by the agents throughout (left or right) were also randomized, with the constraint that an equal number of 'Left' and 'Right' decisions were made to avoid participants developing a location bias. The four fictitious agents were represented by abstract avatars that only differed in their color, the assignment of which was counterbalanced across participants.

Procedure The experiment took on average 55minutes, including a break halfway through the experiment. The break was used to decrease the likelihood that participants would carry over inferences about avatars previously seen at a given position over to the new avatars at the same position. A minimum 3 minute break was enforced, but participants were allowed to take as long as 30minutes.

Materials The alien images were taken from the symmetric Greebles dataset and are presented courtesy of Michael J. Tarr, Carnegie Mellon University, http://www.tarrlab.org/.

Predictions The main prediction concerned the binary decisions of participants, specifically how their choices should vary as a function of the confidence difference between the two advisers' confidence. The null hypothesis was that agents rely solely on explicit differences in confidence when choosing advisers, specifically, on a trial-by-trial basis choosing the answer proposed by the adviser who had a higher level of stated confidence on that trial. Alternatively, agents could make decisions based on differences in the agents' probability of being correct given their stated confidence (recalibrated confidence). For the conditions in which agents had equal accuracy, this results in the predictions in Figure 4. An underconfident agent is more likely to be accurate in our design (and therefore should be chosen) even when they are slightly less confident (>15%) than the calibrated agent (Condition A). To the contrary, an overconfident agent should not be chosen when they are sightly more confident (<15%) than a calibrated agent (Condition B).

Data analysis Participants were removed from the analysis based on two preset criteria. First, participants had to respond above chance on the attention checks. Second, for the main analysis, only participants whose decisions varied with the confidence of the two advisers were included. In order to assess this, a logistic regression was fitted for every participant's bets using the advice of the two agents as predictors. The Aikaike Information Criterion (AIC) was used to compare this model to a random response model. Eight participants were excluded from the analysis based on these criteria (which were both fulfilled).

A Bayesian generalized mixed effects logistic model was fit to the binary choice data (whether the calibrated agent's advice was chosen or not on a given trial), using the difference in the confidence of the advisers (confidence of calibrated agent - confidence of other agent) as a fixed predictor, and participants as a random intercept. The intercept and slope of each participant were assumed to be sampled from Gaussian distributions with unknown mean and standard deviation. The hyperpriors on the population mean and standard deviation were generic weakly informative priors following Gelman, Carlin, Stern, and Rubin (2003). The means were assumed to come from a Student's t distribution and standard deviations were sampled from the Half-normal distribution. A model was fit separately for every condition, resulting in a posterior distribution for the categorization boundary that could be compared to predictions.

To check that participant choices were indeed driven by the difference in the confidence of the two agents, and were not dominated just by just one of the agent's choices, the same logistic model was fitted using both agent's judgements as predictors. The relative weights assigned to the two agents were not statistically different.

We expected that continuous judgements would mirror effects observed in the binary choices, assuming that participants compute their own confidence in each of the two possible answers and choose the one with the highest confidence. Visual exploration of the data revealed that continuous (absolute) confidence judgements were in fact mostly driven by

the confidence of the chosen agent. A priori the expectation is that the confidence of the participant should be influenced by both the confidence of the agent whose advice was taken as well as by the confidence of the other adviser. Intuitively, the higher the confidence of the adviser whose answer was chosen, the higher the confidence of the participant in her answer. Conversely, due to the inherent disagreement in the adviser recommendations, the higher the confidence of the not chosen agent (in the opposite answer), the less confident should the participant be in their selected answer.

In order to explore the influence of the confidence of the chosen adviser and that of the other adviser on the participants' continuous confidence ratings, we regressed participants' confidence ratings on those of the agents and statistically compared the resulting weights.

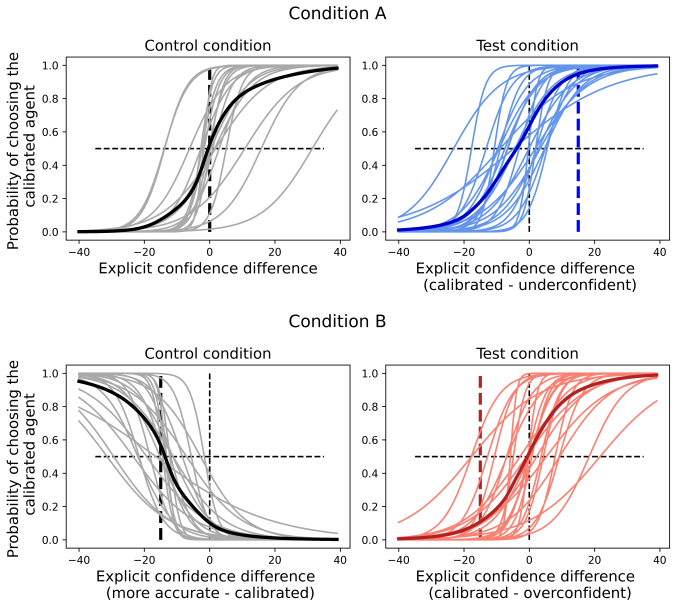

Figure 5: Fitted logistic curves for each participant's adviser choices as a function of explicit confidence differences (calibrated - other agent). The y-axis is the probability of choosing the calibrated adviser. In condition A control block both agents are calibrated, so one agent was randomly chosen as the reference agent for the plotting. Dashed vertical lines represent predictions for the boundaries based on recalibration and bold sigmoids are condition averages.

Results

Adviser choices: Control blocks

The most likely boundary for the control block of condition A, when the two presented agents were in fact identical, was -.08, 95% Highest density interval (HDI):[-2.82, 2.69]. As seen in Figure 5, there was very little variability in the individual boundaries of participants, which closely clustered around zero, $M_{boundary} = 1.34, t(18) = .58, p = .57, BF_{null} = 3.62$.

In the control block of condition B, when the two agents had the same confidence, but differed in accuracy, participants' boundaries shifted, as the more accurate agent was chosen beyond what would be expected from explicit confidence differences. The maximum a posteriori estimate (MAP) for the boundaries was -13.16, 95% HDI [-17.38,-9.55]. The predicted difference (15) was included in the HDI, but zero was not. Individual boundaries showed high consistency, $M_{boundary} = -14.73, t(21) = -8.97, p < .001, BF_{alt} > 10^5$.

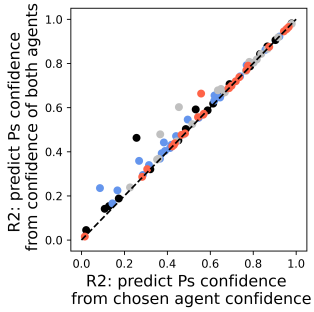

Figure 6: R^2 for predicting individual confidence judgements from the confidence of the potential advisers. Each dot is a participant, all conditions are overlaid (red: cond A test; blue: cond B test; gray: cond A control; black: cond B control).

Adviser choices: Test blocks

Results did not follow predictions in the test blocks. In condition A, the boundary MAP estimate was negative, -3.77, 95% HDI [-6.83, -.85], suggesting that the more confident (and calibrated) agent's advice was used even when they were somewhat less confident than the underconfident agent. Individual boundaries were also predominantly negative, $M_{boundary} = -4.48, t(23) = -2.92, p < .01, BF_{alt} = 6.09$.

There was no discernible pattern at the group level in condition B, as individual boundaries varied widely around zero $M_{boundary} = -.16, t(25) = -.09, p = .93, BF_{null} = 4.81$. The MAP estimate for the boundary was -.43, 95% HDI [-3.22, 2.29].

Confidence judgements

Contrary to predictions, the continuous confidence judgements produced by participants overwhelmingly depended on the confidence of the agent whose advice was taken across all conditions (see Figure 7), and the confidence of the other agent was not meaningfully incorporated in their stated confidence. This is evident from the fact that removing the confidence of the agent whose advice was not taken from the model did not decrease the amount of explained variance for the vast majority of participants (see Figure 6).

Future collaborator preferences

When the two agents were identical (A), 2AFC preferences were random $prop = .47, z = -.36, p = .71, BF_{null} = 2.30$.

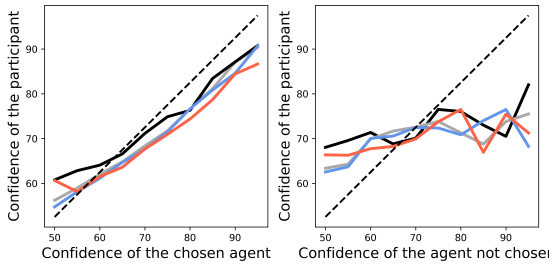

Figure 7: The confidence of participants closely tracks the confidence of the adviser they selected (Left), but is not related to the confidence of the adviser who was not picked (Right). Colors indicate experimental conditions.

Participants preferred to collaborate with the more accurate agent (B), $prop = .90, z = 7.30, p < .001, BF_{alt} = 3279.49$. In the test block of condition A, when deciding between an underconfident agent and a calibrated one with equal accuracy, the majority of the participants chose the more confident (and calibrated) agent, $prop = .77, z = 3.45, p < .001, BF_{alt} = 16.56$. In condition B, participants had a very slight non-significant preference for the overconfident (miscalibrated) agent over the calibrated one, $prop = .60, z = 1.12, p = .26, BF_{null} = 1.48$.

Discussion

Contrary to our predictions, participants did not make fine-grained optimal decisions about whose advice to take based on differences in the advisers' true probability of being correct on a given trial. This was not due to a failure to understand the task since performance in the control blocks conformed to expectations. The failure to adapt decisions to the adviser profile is notable given that recalibration in this case was particularly easy (only a constant bias adjustment).

In condition A, when comparing advisers who differed in their confidence, but were matched for accuracy, participants were unduly influenced by the more confident adviser. This happened even though participants had observed the potential advisers over an extended number of trials and were presented with summary statistics to ensure that differences would be salient.

Results from condition A alone could be interpreted as lending support to the confidence heuristic. However, based on the lack of a similar pattern in condition B, we suggest that calibration was a mediating factor. Specifically, a more confident agent held more sway on participants when it was calibrated, but not when it was overconfident. It should be noted that this modulation of calibration was not actually beneficial in our task. The differential outcome in the two conditions also suggests that the magnitude of the confidence manipulation was sufficient for participants to pick up on calibration differences, although, further confirmation with larger differences is needed.

However, we need to exert caution in the interpretation of the condition A and B differences given the null pattern in condition B was the consequence of large inter-individual variability (that we could not explain based on the measure of task attentiveness). Replication and extension of the experiment with additional adviser profiles is needed before we can conclusively reject the alternative that participants were using a confidence heuristic meanwhile having a noisy perception of accuracy differences. The source of the inter-individual variability is an interesting further direction in itself, especially since they did not relate significantly to differences in accuracy estimation performance.

Further, while the experimental design used a linear mapping of confidence values to the visually presented scale, it is possible that participants assumed a non-linear mapping. Prior training of participants on the confidence judgement task may help address concerns about mapping judgements to the scale in future studies. Related to this point, we also assumed that participants themselves believed implicitly that the two agents used the same mapping from their confidence ratings to the visually presented scale. This is unlikely to be true of real advisers(Bang et al., n.d.), and it is an open question whether observers make this assumption or not.

Importantly, there was no indication in our experiment that either of the advisers had a motive to (or would incur any benefit from) strategic manipulations of their confidence. It is possible that in situations where the two advisers are competing (with each other or for the influence of the participant), people would exert more vigilance and results would more closely match our predictions.

Alternatively, it is possible that, especially in competitive settings, instead of engaging in effortful trial-by-trial recalibration, participants would use calibration information to build global adviser preferences as a function of their perceived trustworthiness and simply discount new information provided by untrustworthy advisers. In line with this, previous work has shown that people are more likely to take the advice of advisers they trust (Sniezek & Van Swol, 2001). In the current experiment, participants assigned equal weights to the information provided by the two advisers, even when there was an accuracy difference between the advisers, suggesting participants exploited both sources of information.

The dissociation between the way in which participants made decisions about whose advice to take and how they computed their confidence in their judgements merits further attention. In the current design, there is some ambiguity between participants truly reporting confidence in their decision or confidence in the adviser. If it is indeed the case that confidence of advice takers is entirely determined by the confidence of the person whose advice was selected, and more confident advisers are generally preferred (unless they are blatantly wrong), this can further amplify overconfidence as information is being circulated in social networks.

References

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. doi: 10.1126/science.1185718

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., … Summerfield, C. (n.d.). Confidence matching in group decision-making. , *1*(6), 0117. Retrieved from https://doi.org/10.1038/s41562-017-0117 doi: 10.1038/s41562-017-0117

Birch, S. A. J., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others non-verbal cues to credibility. *Developmental Science*, *13*(2), 363–369. doi: 10.1111/j.1467-7687.2009.00906.x

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. 2nd ed. CRC Press, London. MR2027492.

Kurvers, R. H., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., & Bahrami, B. (2021). Strategic disinformation outperforms honesty in competition for social influence. *iScience*, *24*(12), 103505. doi: https://doi.org/10.1016/j.isci.2021.103505

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., … Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *PNAS*, *112*(12), 3835–3840. doi: 10.1073/pnas.1421692112

Martí, L., Mollica, F., Piantadosi, S., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind*, *2*(2), 47–60. doi: 10.1162/opmia00017

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965. doi: 10.1037/xge0000180

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57. doi: 10.1002/bdm.460

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*(2), 246–255. doi: 10.1016/j.obhdp.2013.02.001

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288-307. doi: https://doi.org/10.1006/obhd.2000.2926

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792. doi: 10.1016/j.concog.2010.12.011

Sperber, D., Clament, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359-393. doi: https://doi.org/10.1111/j.1468-0017.2010.01394.x

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50. doi: 10.1111/j.1467-9280.2007.01847.x

Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, *16*(1), 97–113. doi: 10.1016/0167-4870(94)00032-6